The use of supermarket loyalty card transaction records as a source of population dietary information

Victoria Louise Jenneson

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

The University of Leeds
Leeds Institute for Data Analytics
School of Geography

February 2022

Intellectual Property and Publication Statements

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

Chapter 3 entitled ‘A systematic review of automated electronic supermarket sales data for population dietary surveillance’ is an exact copy of the journal article:

> Jenneson VL, Pontin F, Greenwood DC, Clarke GP, Morris MA. 2021. A systematic review of supermarket automated electronic sales data for population dietary surveillance. *Nutrition Reviews*, Nuab089

VJ was lead author and reviewer, designed the protocol, developed the initial data extraction form, reviewed papers against inclusion criteria, and carried out data extraction, thematic analysis and write-up. FP was second reviewer, co-designed the data extraction form, and conducted second review of papers against the inclusion criteria. DCG, GPC and MAM supervised the worked and provided advice on the design of the review protocol and conduct of the review. They gave substantial support and comments on drafts of the review. Additionally, MAM was the third reviewer and helped with resolution of conflicts where a decision could not be reached between reviewer 1 and reviewer 2.

Chapter 4 entitled ‘Exploring the geographic variation in fruit and vegetable purchasing behaviour using supermarket transaction data’ is an exact copy of the journal article:

> Jenneson V, Clarke GP, Greenwood DC, Shute B, Tempest B, Rains T, Morris MA. 2022. Exploring the Geographic Variation in Fruit and Vegetable Purchasing Behaviour Using Supermarket Transaction Data. *Nutrients.* **14**(1), 177

VJ conducted the analysis and write-up of results. GPC, DCG and MAM supervised the work and provided support on the design and conduct of the analysis, as well as feedback on drafts of the paper. GPC provided expertise on spatial analysis, DCG provided expertise on statistical analyses and interpretation. BS, BT and TR supported the work on behalf of Sainsbury's. They facilitated data sharing and commented on drafts of the paper.

Chapter 5 entitled 'Supermarket Transaction Records In Dietary Evaluation – The STRIDE study: validation against self-reported dietary intake' is an exact copy of the journal article:

> Jenneson V, Greenwood DC, Clarke GP, Rains T, Tempest B, Shute B, Morris MA. 2022 [In Preparation]. Supermarket Transaction Records In Dietary Evaluation – The STRIDE study: validation against self-reported dietary intake'. *Public Health Nutrition.*

VJ designed and implemented the STRIDE study protocol, including recruitment, data collection, analysis and write-up of results. GPC, DCG and MAM supervised the work and provided guidance on the design and conduct of the study and comments on drafts of the paper. DCG provided expertise on statistical analysis and study design, and MAM provided expertise on study design and data storage protocols, including advice on working in partnership with a commercial organisation and sign-off of data-sharing agreements.

Acknowledgements

I am extremely grateful for the support I have received from across the University of Leeds and beyond, without which the work in this thesis would not have been possible.

Firstly, I would like to think my academic supervisors, Michelle Morris, Darren Greenwood and Graham Clarke for their kindness, patience and support (both intellectually and personally). Their commitment not only to the project, but to my development as a researcher has contributed to a fantastic experience and allowed my confidence to grow immensely over the last four years. I would also like to acknowledge my Research Support Group, Hannah Ensaff, Nik Lomax and Paul Norman for their feedback during the earlier stages of this work. Thanks also go to Francesca Pontin for her input as second reviewer for the systematic review, and to Stephen Clark for being so generous with sharing his expertise in coding and data preparation, and for helping me to solve many a problem.

This work was funded by the Economic and Social Research Council (ESRC) Centre for Doctoral Training in Data Analytics and Society. I am grateful to the CDT for the excellent foundation in data science training and for the nurturing research community. In particular, I would like to thank Hayley Irving, Claudia Rogers and Eleri Pound. In addition, I would like to acknowledge the many teams within the University who have supported this work: The LIDA Data Analytics Team (particularly Sean Tuck and Adam Keeley for their assistance with data access), The LIDA communications team (particularly Robyn Naisbitt for her help to launch the STRIDE study); and colleagues in the admin, legal and ethical review teams for enabling this to work happen. I would also like to thank the Scottish Collaborative Group team at the University of Aberdeen for providing access to and support with their online Food Frequency Questionnaire.

My experience over the last four years was shaped by the amazing friendship of my fellow PhD students, whose humour, pep-talks, and emotional support helped to keep me motivated throughout. To Francesca Pontin, Rachel Oldroyd, Charlotte Sturley, Amanda Otley and Ryan Urquhart, it was a pleasure to share this with you.

This work could not have been accomplished without the trusted relationships built with colleagues at Sainsbury’s, including the Data Governance and Information Security Teams, Consumer Insights Team, and Nectar Team. In particular, thanks go to Nilani Sritharan, Bethan Tempest and Becky Shute in

the Nutrition team, for championing my work within the business and for welcoming me as an extension to their team, and to Tim Rains for the generosity of his time and support with customer sampling.

Thanks also go to my fantastic support network outside of the University. To Catherine Drake, for her many years of friendship and always knowing just when I needed to talk. To my family, Wayne Litherland, Sally Litherland, Emma Boden and Daniel Litherland, for always believing in me, even when I don't believe in myself. Finally, to my husband Luke, thank you for being my rock throughout, for making sure I left my desk from time to time for a dog walk and a chat, for keeping the fridge stocked and making my tea (I don't remember the last time I cooked a meal!), and for holding onto your own dreams while I pursued mine. I couldn't have done it without you, and now it's your turn to shine.

Abstract

Recognising the need to intervene on the food system to reduce obesity and associated non-communicable disease, governments have begun to apply legislative sanctions for the sale and advertising of unhealthy foods. Suitable metrics are therefore critical to monitor population dietary behaviours, supporting the design and evaluation of policy interventions.

National dietary surveys reveal trends, but their coverage is limited, leaving hard-to-reach low-income and minority ethnic groups largely under-represented by population dietary statistics. Supermarket loyalty card transactions offer simple, consistent, scalable dietary metrics to complement existing approaches.

In partnership with Sainsbury's, this thesis explores the utility of supermarket loyalty card transaction data as a novel tool for population dietary monitoring. This begins with presenting results from a systematic literature review, which identifies research gaps addressed in the remainder of the thesis. Using transactions for around 50,000 loyalty card customers, in Leeds, I present a unique spatial exploration of dietary purchases at the neighbourhood level. Spatial clustering observed in fruit and vegetable purchasing patterns supports the association between deprivation and poor dietary quality, as well as identifying areas which oppose this trend. This demonstrates the capacity of transactions data to contribute to hypothesis generation in ecological research and the targeting of local dietary policy strategies.

Through conduct of the STRIDE study, this thesis for the first time quantifies agreement with an online food frequency questionnaire for 686 participants, adding to knowledge of the validity of transactions as a nutrient-level dietary metric. Purchases demonstrate good agreement with intake for energy-adjusted metrics, making them a good marker of individual-level dietary composition and diet quality. Yet, the evidence for agreement with absolute measures is less clear, with variation by household size and loyalty.

Identification and characterisation of appropriate customer samples will be key for the generalisability of absolute purchase metrics as a population-level dietary proxy.

Table of Contents

Intellectual Property and Publication Statements ii
Acknowledgements...iv
Abstract...vi
Table of Contents vii
List of Tables xi
List of Figures xii
List of Abbreviations..xiv
Chapter 1 Background and rationale...17
1.1 An overview of dietary assessment.......................................18
1.2 The role for technology in dietary assessment 20
1.3 Reframing dietary assessment...22
1.4 Food purchase data for population dietary assessment........26
1.5 Thesis aim...30
1.5.1 Objectives 30
1.6 Background chapter summary 31
Chapter 2 Data Preparation and Methods 32
2.1 Retail transaction data (2016 sample)..................................32
2.2 The STRIDE Study (Supermarket Transaction Records In Dietary Evaluation)...36
2.2.1 Designing the STRIDE study......................................37
2.2.1.1 Data Governance 37
2.2.2 STRIDE study recruitment..38
2.2.3 STRIDE data collection 46
2.2.3.1 Baseline questionnaire....................................46
2.2.3.2 Dietary intake assessment 47
2.2.3.3 Food purchase data 50
2.2.4 Development of a bespoke nutrient composition database 50
2.2.5 Dietary estimation 52
2.2.5.1 Daily intake estimates 52
2.2.5.2 Daily purchase estimates................................53
2.2.6 Data linkage 56
2.2.7 Impacts of COVID-19 58
2.3 Data and methods chapter summary 58

References 60

Chapter 3 A systematic review of automated electronic supermarket sales data for population dietary surveillance .68

Abstract 68

3.1 Introduction..69

3.2 Methods..72

3.2.1 Search strategy 72

3.2.2 Study selection and data extraction..73

3.2.3 Quality assessment 75

3.2.4 Data synthesis..75

3.3 Results 76

3.3.1 Search results 76

3.3.2 Study characteristics 76

3.3.3 Risk of bias..79

3.3.4 Study aims..79

3.3.4.1 Evaluating intervention effectiveness 80

3.3.4.2 Dietary surveillance 81

3.3.5 Dietary assessment..84

3.3.5.1 Representativeness of total household purchasing..84

3.3.5.2 Representativeness of individual food consumption..85

3.3.6 Sources of nutrient data 86

3.3.7 Outcomes..89

3.4 Discussion..91

3.4.1 Future research priorities 96

3.5 Conclusion..97

3.6 Acknowledgements 97

3.6.1 Funding 97

3.6.2 Declaration of interest 97

3.7 Supplememtary materials..98

References..139

Chapter 4 Exploring the geographic variation in fruit and vegetable purchasing behaviour using supermarket transaction data..148

Abstract 148

4.1 Introduction..148

4.2 Methods 151
4.2.1 Study sample 151
4.2.2 Study region 152
4.2.3 Transaction data 154
4.2.4 Estimating fruit and vegetable purchases.................155
4.2.5 Analysis...155
4.3 Results 156
4.3.1 Customer characteristics.....................................156
4.3.2 Fruit and vegetable purchases in Leeds..................157
4.3.3 Neighbourhood characteristics...............................160
4.3.4 Spatial patterns in fruit and vegetable purchasing....164
4.3.5 Linear regression 165
4.4 Discussion..168
4.4.1 Policy relevance 174
4.4.2 Future directions 175
4.5 Conclusions..175
4.6 Supplementary materials...175
4.7 Additional information...177
References...178
Chapter 5 Supermarket Transaction Records In Dietary Evaluation – The STRIDE study: validation against self-reported dietary intake...184
Abstract...184
5.1 Introduction 185
5.2 Methods 187
5.2.1 Study design 187
5.2.2 Participant sampling and recruitment 187
5.2.3 Data collection...189
5.2.4 Nutrient estimates 190
5.2.5 Statistical analysis...191
5.3 Results 192
5.4 Discussion...205
5.4.1 Future research avenues 208
5.5 Conclusion 209
References..215

Chapter 6 Discussion..219
6.1 Overview 219
6.1.1 Key Findings...220
6.2 Nutrition data metrics 221
6.2.1 Product coverage 222
6.2.2 Nutrient coverage and accuracy................................224
6.2.3 Categorisation approaches..225
6.3 Spatial granularity..227
6.3.1 Contribution to ecological research 227
6.4 Population coverage..229
6.5 Representativeness of diet...230
6.5.1 Coverage of total purchases.....................................233
6.5.2 Extrapolation to the individual-level...........................235
6.5.3 Bias in consumption estimation................................236
6.5.4 Agreement by dietary metric.....................................237
6.6 The importance of partnerships..238
6.7 Impacts of the COVID-19 pandemic.....................................240
6.8 Future research directions..240
6.9 Conclusions...241
References..243
Appendix A Systematic Review supporting documents..............250
A.1 Systematic review protocol (as published on PROSPERO) 250
Appendix B STRIDE supporting documents.................................256
B.1 STRIDE data management plan 256
B.2 STRIDE ethical review form...262
B.3 STRIDE participant information...275
B.4 STRIDE informed consent form 280
B.5 STRIDE baseline questionnaire..281
B.6 SCG-FFQ Fat Coding Sheet for researchers.....................287

List of Tables

Table 1.1 Dietary assessment methods for each of the 5 stages of the Framework for Evaluating Diet (FED)................................25

Table 2.1 STRIDE expected recruitment figures (pre-pilot)..........40

Table 2.2 Revised expected sign-up and completion figures for STRIDE cohorts..46

Table 2.3 UK recommended daily calorie intakes by age and gender 55

Table 3.1 PICOS criteria for inclusion and exclusion of studies..73

Table 3.2 Summary of included paper characteristics..................78

Table 4.1 Food categories in the transaction database used for sampling 152

Table 4.2 Coverage of study sample by demographic group, in relation to Leeds and UK..158

Table 4.3 Overview of variables at Lower Super Output Area level 161

Table 4.4 Results of OLS regression predicting household fruit and vegetable purchasing (portions/day) 166

Table 5.1 UK recommended daily calorie intakes by age and gender 191

Table 5.2 STRIDE participant recruitment summary...................192

Table 5.3 STRIDE participant characteristics for all cohorts (including pilot) combined 194

Table 5.4 Absolute nutrient estimates from purchase records and food frequency questionnaires (n = 686) 196

Table 5.5 Energy-adjusted nutrient estimates from purchase records and food frequency questionnaires (n = 686).........197

Table 5.6 Mean difference and limits of agreement between purchase and intake for energy (kcal)...................................201

Table 5.7 Mean difference and limits of agreement between individual-level purchase estimates and intake for macronutrients and sodium .... Error! Bookmark not defined.

Table 5.8 Mean difference and limits of agreement for energy-adjusted purchase and intake for macronutrients and sodium 202

List of Figures

Figure 1.1 Framework for Evaluating Diet (FED). 23

Figure 2.1 Secondary retail data relations diagram (2016 sample). 35

Figure 2.2 Participant journey flow diagram for STRIDE pilot study..41

Figure 2.3 STRIDE pilot baseline survey analytics........................42

Figure 2.4 STRIDE invitation test email..43

Figure 2.5 STRIDE invitation email terms and links to privacy notices..43

Figure 2.6 STRIDE baseline survey and consent form homepage 44

Figure 2.7 Screenshot of STRIDE study webpage containing participant information 44

Figure 2.8 Screenshot of STRIDE study webpage – More information headings..45

Figure 2.9 Screenshot of Scottish Collaborative Group online Food Frequency Questionnaire (SCG-FFQ) landing page.....48

Figure 2.10 Screenshot of SCG-FFQ Page 1..................................48

Figure 2.11 STRIDE data relations diagram 57

Figure 3.1 Study selection PRISMA flow chart for a review of the use of electronic sales records in population dietary surveillance – (June 2020)...77

Figure 4.1 (a) Index of Multiple Deprivation decile by Lower Super Output Area in Leeds. (b) Output Area Classification by Output Area in Leeds. 154

Figure 4.2 Fruit and vegetable purchasing in Leeds by Lower Super Output Area: mean daily portion per household. 164

Figure 4.3 Local Moran's I for daily fruit and vegetable portions per household...165

Figure 4.4 Map of residuals from linear regression model.........168

Figure 5.1 STRIDE study design 187

Figure 5.2 STRIDE participant journey flow diagram..................189

Figure 5.3 Bland-Altman plots for ratio of energy measurements plotted against their average..........Error! Bookmark not defined.

Figure 5.4 Bland-Altman plots for ratio of energy purchased (individual-level)/energy intake, plotted against their average, by household size. ....Error! Bookmark not defined.

Figure 5.5 Bland-Altman plots for ratio of energy purchased (individual-level)/energy intake, plotted against their average, by self-reported customer loyalty.. Error! Bookmark not defined.

Figure 5.6 Bland-Altman plots for ratio of nutrient purchased (individual-level)/nutrient intake, plotted against their average, by nutrient. .... Error! Bookmark not defined.

Figure 5.7 Bland-Altman plots for ratio of energy-adjusted nutrient purchased (individual-level)/energy-adjusted nutrient intake, plotted against their average, by nutrient. 204

Figure 6.1 Framework for Evaluating Diet (FED) showing indicative proportion of food from a single retailer, outlined in yellow. 232

List of Abbreviations

BMI Body Mass Index

BOP Back of Package

CDRC Consumer Data Research Centre

CoFID Composition of Foods Integrated Dataset

COICOP Classification Of Individual Consumption by Purpose

CSR Corporate Social Responsibility

DAT Data Analytics Team

DLA Data Licence Agreement

DMP Data Management Plan

EAN European Article Number

EPOS Electronic Point of Sale

ESRC Economic and Social Research Council

FBDGs Food-based Dietary Guidelines

FCDB Food Composition Database

FED Framework for Evaluating Diet

FFQ Food Frequency Questionnaire

FFS Family Food Survey

FV Fruit and Vegetables

GTIN Global Trade Item Number

GWR Geographically Weighted Regression

HFSS High in Fat Salt and Sugar

IMD Index of Multiple Deprivation

LASER Leeds Analytic Secure Environment for Research

LCFS Living Costs and Food Survey

LIDA Leeds Institute for Data Analytics

LoA Limits of Agreement

LSOA Lower Super Output Area

NDNS National Diet and Nutrition Survey

OA Output Area

OAC Output Area Classification

ONS Office for National Statistics

PEN Penetration Testing

PHE Public Health England

PNCD Product Nutrient Composition Database

PRISMA Preferred Reporting Items for Systematic reviews and Meta-Analysis

RFM Recency, Frequency, Monetary value

SCG Scottish Collaborative Group

SDIL Soft Drinks Industry Levy

SKU Stock Keeping Unit

STRIDE Supermarket Transaction Records In Dietary Evaluation

WHO World Health Organisation

Part I

Introduction

Chapter 1
Background and rationale

In this first chapter, I will set out why population-level dietary monitoring is important and briefly describe the current state of play for how dietary assessment is undertaken. I then go on to describe some of the challenges faced in the development of dietary assessment methods and how harnessing technology, including electronic supermarket transactions, may help to overcome these. I next explain how a shift in the way we consider what constitutes our diet opens up the possibilities of a suite of dietary assessment methods with the capacity to capture different aspects of population diet. By rejecting the notion of a single ‘gold standard’ dietary assessment method, I propose a theoretical framework for diet which conceptualises the movement of foods and nutrients from our food environment through to our bodies. Under this framework, there is a clear place for supermarket transaction records as a means of capturing purchasing decisions and the foods which enter the household, which can complement more established dietary assessment techniques. Finally, I give a brief review of the use of transaction records in dietary research, before concluding the chapter with the nine objectives set out by this thesis to achieve its overall aim:

Aim: To understand and evaluate the contribution that supermarket loyalty card transaction records can make to population dietary monitoring, both on a national and a small-area scale.

In the subsequent chapters, I will outline the data and methods employed to meet these objectives (Chapter 2), providing more detail to that which is provided in the three papers which make up this alternative format thesis. Chapter 3 in this thesis (paper 1) systematically reviews the use of automated electronically captured supermarket transaction records for dietary surveillance, bringing the reader up to date with the progress in the field and research gaps upon which the next two papers build. Chapter 4 (paper 2) presents an application of supermarket transaction data to explore small area geographic variation in dietary purchase behaviours, using fruit and vegetable purchases as a proxy for healthy diets, in the city of Leeds, England. In Chapter 5 (paper 3), I present a paper describing the results of the STRIDE validation study (Supermarket Transaction Records In Dietary Evaluation), in which the statistical agreement between dietary estimates from supermarket loyalty card transaction records and self-reported intake is quantified. The final chapter of the thesis (Chapter 6), demonstrates how the overall aim is achieved through meeting each of the nine objectives. In pulling together the findings from each of the three papers which make up the thesis, I discuss the unique contribution of this work to the field of population dietary assessment, outlining future research avenues to harness the capacity of supermarket loyalty card transactions.

1.1 An overview of dietary assessment

Poor dietary quality is a major factor contributing to the rising rates of obesity and associated comorbidities in the UK. Diets lacking in fruits, vegetables and wholegrains, and high in salt, sugar and saturated fats lead to an overconsumption of calories and are associated with an increased risk of Type 2 Diabetes (Neuenschwander et al., 2019), cardiovascular disease (Casas et al., 2018), and some types of cancer (WCRF, 2018), among other illnesses. Diet-related ill health is estimated to contribute up to half of total global disease (World Health Organization., 2003). Conditions related to high BMI alone have been projected to contribute around 8% (£18 billion annually) of the UK's total healthcare spend between 2020 and 2050 (OECD, 2019; National Food Strategy., 2021). Dietary surveillance is therefore important as it enables epidemiologists to better understand the interactions between diet and health, and permits the observation of population-level trends and development of health improvement strategies.

Biologically, a person's nutritional status (i.e. the amount of nutrients present in one's body) is linked to their health status. The nutrients and other non-nutritive food compounds in our bodies can influence the efficiency of our digestive system and make-up of our gut microflora (Valdes et al., 2018), the level of chemicals in our bloodstreams (Russell et al., 2016), our cognitive function (Gutierrez et al., 2021), and many other biological processes. Nutritional biomarkers are an important tool for objectively monitoring nutritional status, either directly via concentration biomarkers in blood or tissue, or indirectly through recovery biomarkers or predictive biomarkers found in excretory products (Dietary Assessment Primer., 2022). However, not all dietary components have suitable biomarkers. Furthermore, we must acknowledge that people eat foods rather than nutrients. Therefore, if we wish to intervene to improve nutritional status it is important to understand the upstream dietary behaviours which lead to nutritional status. Thus, we concern ourselves with dietary assessment.

Although in common parlance the word 'diet' is used to describe restrictive eating practices usually for weight loss purposes. In the context of nutrition research 'diet' is commonly considered to mean 'the foods that someone usually consumes'. Dietary assessment provides a useful proxy for nutritional status. After all, food (and its constituent nutrients) must be eaten in order to interact with the body. Therefore, if we know what someone is eating, we may be able to hypothesise as to their risk of developing a given disease and advise dietary improvements where appropriate. Comparative to biomarkers, dietary assessment is typically less invasive, quicker, cheaper and easier to administer. Trends in dietary intake at the population level are therefore a surrogate for population nutritional status, enabling us to develop new hypotheses, generate population dietary guidance and public health policy. Due to the important role that dietary surveillance plays, research into dietary assessment methods therefore concerns itself with developing better ways to record what someone habitually puts in their mouth.

However, measuring what someone eats is inherently difficult. It is estimated that each of us make around 200 food-related decisions each day, most of which are unconscious, contributing to under-estimation of intake (Wansink, 2010). To capture food consumption objectively would mean an invasion of privacy likely to put the majority of study participants off, such as the installation of cameras in the home which would capture more than just eating behaviours. Researchers therefore depend largely on self-reported records of intake such as 24-hour recalls, food diaries, and Food Frequency Questionnaires (FFQs). Yet, eating and drinking are such routine occurrences in everyday life that they often happen subconsciously. The ability to accurately recall exactly what was consumed (and the exact quantities) even for the previous day is therefore problematic and leads to recall bias.

Food is also bound up in socio-cultural norms defined by our experiences with the world. This makes us more likely to under-report consumption of foods, particularly those we understand to be unhealthy (such as those high in fat, salt and sugar) and over-report those we understand to be healthy (such as vegetables). This is known as social desirability bias (Hebert et al., 1995). Reporting biases also differ according to our personal characteristics, such as gender (Hebert et al., 1997; Hebert et al., 1995) and BMI (Wehling and Lusher, 2019). While prospective food records such as diaries remove the need for even short-term memory, the very process of recording our consumption has a tendency to temporarily modify our behaviours such that they are perceived as more socially desirable, a type of bias known as the Hawthorne effect (McCambridge et al., 2014).

Aside from their subjectivity, self-reported dietary intake assessment methods also suffer from self-selection biases, even if randomisation techniques are used to maximise sample representativeness (Bonevski et al., 2014; Rehm et al., 2021). As survey-based research attracts participants with an interest in the subject, dietary research participants tend to: exhibit more favourable eating habits; be healthier; more highly educated; less deprived; or more highly motivated towards health change due to their underlying health status (Rehm et al., 2021; Bonevski et al., 2014). Self-reported dietary assessment methods also carry a substantial burden for participants (especially in the case of weighed food diaries), and for researchers with regards to data input and coding to nutrient databases. These burdens translate to high research costs and impact the scalability and temporality of data collection; after all, participants are only willing to record their food consumption for so long. As a result, dietary surveys such as the UK's National Diet and Nutrition Survey (NDNS), tend to be cross-sectional, collecting data over just four days for different participants each year, and are limited to relatively small yet nationally representative samples.

1.2 The role for technology in dietary assessment

Advancements in technology have enabled the development of new online versions of established self-reported dietary assessment methods, such as Intake24 (Simpson et al., 2017) and myfood24 (Carter et al., 2015) which permit both retrospective dietary recall and prospective diary-style food intake data capture. Online dietary assessment methods benefit from scalability due to their inexpensive roll out and automated entry coding, as well as reductions in researcher coding errors. Recently, the NDNS has moved to using an online dietary assessment using Intake24 (Simpson et al., 2017; PHE, 2021) in place of paper food diaries. Despite these advantages, a number of technical and usability issues have been identified with the use of online methods (PHE, 2021) and they remain subject to self-report and selection biases which limit their accuracy and generalisability.

The reliance on self-reported dietary intake assessment has contributed to a perceived lack of rigour in nutrition research. Combined with the observational nature of nutritional studies, it is considered that poor quality evidence has contributed to insufficient policy success and loss of public trust (de la Hunty et al., 2021; Theis and White, 2021). As a result, a recent MRC Review of Nutrition and Human Health Research described nutrition as a field in crisis (MRC, 2017). The nutrition research community in the UK responded by convening a two-day Nutrition Research Partnership (NRP) workshop to discuss the need for change in the field (de la Hunty et al., 2021). The workshop concluded that there is a collective desire to seek scalable objective dietary measures which are easy and affordable to implement and acceptable to all sectors of the population. Furthermore, there was a recognition that no single dietary measure can tell us all we need to know and that the wide variety of questions asked of diet and health researchers require different types of data to answer them. The NRP workshop (de la Hunty et al., 2021) and a WHO workshop from the same year (WHO, 2021) both called for currently under-utilised technologies, including purchase data, to be harnessed to procure a broader suite of methods. This breadth of data sources might then be combined in a complementary manner to offer a more complete picture of dietary habits than is possible with any single method.

One technological solution to objectively measure dietary intake is the use of cameras. Body-worn or in-home cameras continually capture images which record the foods consumed without the need for recall or reporting by participants (Gemming et al., 2015; Qiu et al., 2021; Chen et al., 2021). However, recognition of foods and their composition remains a challenge for camera technology and is reliant on high image quality, especially for composite foods such as casseroles which contain numerous components (Gemming et al., 2015). Quantification poses another challenge, particularly in 3D space, with many methods requiring the presence of a reference object or standard plates as a reference for size in images (Gemming et al., 2015). To obtain optimal accuracy, it is recommended that image-capture techniques be accompanied by additional dietary survey information (Gemming et al., 2015). While it is possible that advancements in technology will eventually overcome these issues (Qiu et al., 2021), a substantial hurdle remains around privacy. With the presence of cameras unlikely to appeal to many, the effect of selection bias is likely to be even stronger than for surveys. This, combined with the costs of camera technology would surely limit the sample sizes and data coverage period practicable with the method. Furthermore, there are ethical issues around the capture and management of non-food related images which must be overcome.

Technology may also contribute through the repurposing of secondary big data sources for dietary research. A review of the obesity data landscape in line with nodes of the Foresight obesity system map (Morris et al., 2018) highlighted opportunities to harness commercial big-data sources in the dietary space, including commercial surveys, retail sales, and supermarket loyalty card data. With the advantages of objectivity, scale, granularity and low burden, secondary big data sources merit exploration into their validity for use in population dietary surveillance. The use of supermarket loyalty card data for population-level dietary monitoring will be the focus of this thesis, as described in full in the aim and objectives outlined at the end of this chapter (section 1.5). In the next section (1.3), I propose a reframing of how we consider dietary assessment, which opens up opportunities for insight from currently under-utilised data sources, such as loyalty card purchase records.

1.3 Reframing dietary assessment

Dietary assessment is typically focused on measuring usual food consumption at the individual-level as a marker of nutritional status. By measuring food consumption in enough individuals and taking aggregated statistics, we can understand the diet of a population. The term 'diet' has thus become synonymous with intake in the nutrition research field. Yet, in other contexts, the word 'diet' takes on a broader meaning. The Oxford English Dictionary defines 'diet' as *"(the) Customary course of living as to food: way of feeding"* (Oxford English Dictionary., 2021), while Merriam-Webster (2022) describes its medical definition as *"food and drink regularly provided or consumed"*. These definitions expand the concept of diet beyond just that which enters the mouth, and enable us to consider diet in terms of its related behaviours situated in aspects of way of life and food provision. Based on this broader definition of diet as a reflection of 'food-related habits', this thesis proposes a re-conceptualisation of diet as a flow of foods (and their constituent nutrients) through a series of consecutive stages which encompass, but are not limited to, dietary intake. This concept offers up a wider range of opportunities to learn about the food-related behaviours which determine the flow of foods from the consumer food environment through household purchases and individual consumption to biological markers of nutritional status. This can be visualised as the theoretical Framework for Evaluating Dietary (FED) presented in Figure 1.1.

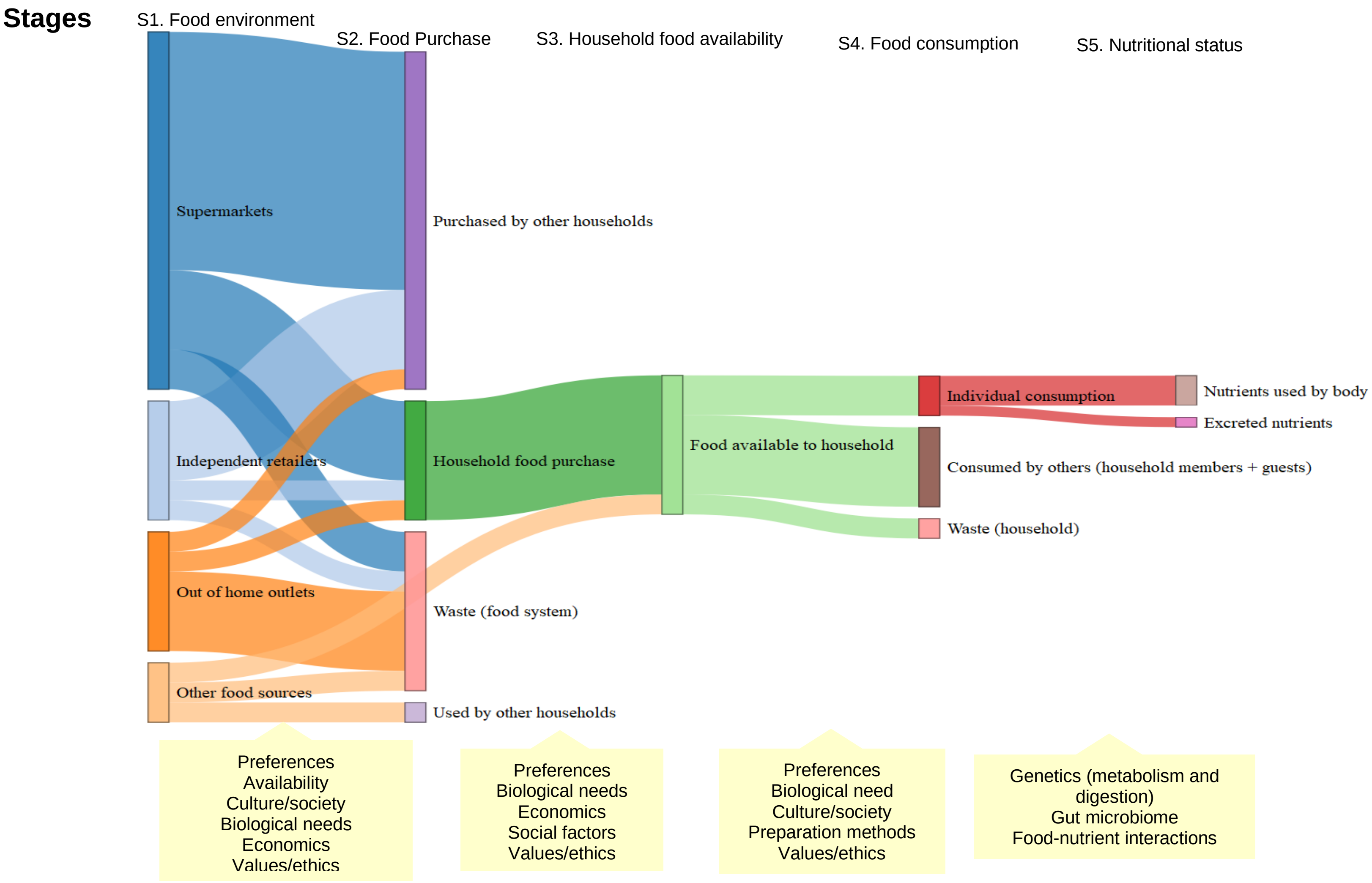


Figure 1.1 Framework for Evaluating Diet (FED).

Paths depict an indicative flow of foods from the consumer food environment through household purchases and individual consumption to biological markers of nutritional status. The thickness of paths is illustrative only and does not represent actual volumes (volumes will vary by person and setting).

The FED diagram depicts the five chronological stages which make up the diet (under our broad food-related habits definition), as a series of coloured vertical bars (Figure 1.1). From left to right across the page, the stages are: S1) the availability of food within the environment around us; S2) the foods we purchase; S3) household food availability; S4) the foods we consume; and S5) our nutritional status. Each stage represents a point at which diet may be measured, offering a different perspective of diet-related habits: at the environment level (S1); at the household level (S2 and S3); and at the individual level (S4 and S5). The transition of foods through each stage of the framework is depicted by paths flowing from left to right (Figure 1.1), the colour of which represents the food's origin. The thickness of these transitional paths is roughly representative of the volume of food and nutrients flowing from the nodes at each stage. However, as the volume will differ between individuals, households and settings, the thickness of paths is for illustrative purposes only and not based on real-world data. For example, as supermarkets are the dominant food source for most UK households, the dark blue paths flowing from the supermarkets portion of the food environment are thicker than the paths which flow from the other food sources.

The foods and their volumes which are permitted to flow from one stage to another are moderated by internal and external constraints which will vary between individuals, indicated by the pale-yellow boxes with arrows at the bottom of the diagram. The thickness of the paths decreases across each stage of the framework as individuals 'choose' which foods they wish to purchase, and consume, and as their unique biological and microbiotal make-up determines which nutrients their body is able to utilise. Not all foods will make it into the individual's body; some are left on the shelf to be purchased by other households, consumed by other members of the household or guests, or contribute to food waste. Not all nutrients which enter the body will be utilised. A person's appetite, time availability, values, and cooking skills are just some of the factors which might determine how the foods inside their kitchen cupboards are transferred to the foods they consume in a single meal or over a given day. In reality, people do not live in a free choice environment (Hawkes et al., 2015). The degree of autonomy an individual is able to exhibit upon these constraints is variable and thus the word 'choice' should be used cautiously. For example, faced with the same supermarket offering (S1 food environment) two individuals with different financial circumstances would not have the same freedoms to purchase the same selection of foods (S2) and would likely come away with very different shopping baskets.

The stages of the FED represent important points for population-level dietary monitoring, while the transitions between them represent opportunities for intervention which may lead to improvements in dietary quality, nutritional status and subsequently health outcomes. Diet may be monitored at each stage of the FED using subjective or objective measures, which I detail further in Table 1.1. While subjective measures at each stage of the framework are well-established, much sought-after objective assessment methods are now emerging thanks to advancements in technology, as previously described. Objective measures avoid the problem of systematic under- and over-reporting associated with self-report. However, given that food consumption is particularly difficult to measure objectively, dietary assessment at S4 (Figure 1.1) has an over-reliance on subjective measures at present. Objective data sources may be more readily collected further upstream in the framework. These include secondary big data sources such as electronically captured supermarket loyalty card purchases, the subject of this thesis, the utility of which is relatively under-explored at present in dietary assessment research.

Table 1.1 Dietary assessment methods for each of the 5 stages of the Framework for Evaluating Diet (FED)

(Stages of the FED as shown in Figure 1.1)

Stage of FED	Subjective measures	Objective measures
S1. Food environment	Self-reported access surveys	Food balance sheets, geocoded store locations, store inventories
S2. Food purchase	Purchase diaries, surveys, market research panel data	Electronic point of sale data, supermarket loyalty cards
S3. Household food availability	Self-reported inventory	Inventory photographs, refrigerator cameras
S4. Food consumption	Food Frequency Questionnaire, 24-hour	Wearable cameras, in-home cameras

	dietary recall, food diaries	
S5. Nutritional status	Self-reported symptoms e.g. fatigue	Nutritional biomarkers

The proposed FED enables consideration of dietary assessment as encompassing a suite of complementary metrics from a variety of sources across different stages of diet-related behaviour. Combining such metrics would offer a more complete picture of population diets and go some way to encapsulating the socio-economic context which drives our food-related habits. The chronological nature of the FED demonstrates that dietary behaviours upstream can influence nutritional status down the line. Therefore, to improve diet-related health outcomes, interventions are required to act at each of the transitions between stages of the framework. It suggests therefore that dietary monitoring at each stage of FED is equally important, with triangulation of datapoints across the whole framework likely to build a more comprehensive picture of population-level dietary habits.

1.4 Food purchase data for population dietary assessment

The body of work presented in this thesis explores the contribution of food purchase monitoring for population dietary assessment (FED S2). Specifically, I focus on the use and validity of supermarket loyalty card transaction records as an objective measure of food purchases. In this section, I outline what makes food purchases an important part of the dietary assessment framework and go on to introduce the unique contribution which loyalty card transactions can offer.

Sitting at the intersection between the food environment and consumption, measuring food purchases offers an opportunity to understand how people interact with their food environment to curate the available food within their household, which is ultimately consumed. As shown in the FED diagram (Figure 1.1), food purchases are constrained by (among other things); economic factors (affordability based on food prices and income); social factors (e.g. a busy working lifestyle); values and ethics (e.g. prioritising health, following a vegan diet or choosing items with fair trade status); individual preferences (e.g. brand affinity or the preference for one flavour over another); and biological needs (e.g. hunger cues at the time of shopping or allergies and intolerances). Food purchase behaviours may therefore be manipulated by changes in food prices, availability, or marketing messages, to influence dietary quality. Furthermore, with the critical importance of the food system on planetary health, food purchases represent an opportune point to intervene and monitor the sustainability impacts of our food choices, whether those foods are ultimately consumed or not (Springmann et al., 2018; Willett et al., 2019).

The UK Government has conducted a survey of food and drink purchases since 1940. Formerly the Wartime Food Survey, the Family Food Survey (FFS) is now an important module of the Living Costs and Food Survey (LCFS), in which a sample of households complete self-reported diaries of all food and drink purchases made over two weeks, including food eaten out, supported by paper till receipts (GOV.UK, 2020a). By tracking food expenditure over time, the FFS provides an understanding of the nation's food-related behaviours, spending power, priorities and attitudes towards food. Our national food-related behaviours typically align with wider social changes. For example, changing roles for women in society and rapid growth in the number of households owning a refrigerator in the 1960s sparked an appetite for convenience foods which persists today (GOV.UK, 2015). Surveillance of dietary purchases also helps us to understand food affordability, contributing to the Consumer Price Index (Office for National Statistics., 2017), and how households in different income groups respond to rising food prices.

However, a limitation of the FFS is the manual burden placed on participants to complete purchase diaries and collect their till receipts, as well as the time required for researchers to code them. As a result, data collection is limited to a two-week period annually for each of the 5,000 participating households, and publications are delayed. At the time of writing, the most recent publication of the FFS is for 2018/19 (published in October 2020) (GOV.UK, 2020a), already two years out of date. Furthermore, a reliance on national food composition databases limits the accuracy of nutrition information used by the LCFS to generate dietary intake estimates.

Market research panels carried out by companies such as Kantar and Nielsen offer an alternative source of food purchase information for a larger sample of participants (Bandy et al., 2019). Panel participants use handheld barcode scanners in their homes to record purchased food items. Supplemented with a book of barcodes for generic unpacked food items (such as fresh vegetables) and a food purchase diary, market research panel data also captures foods purchased for consumption outside of the home. The utility of panel data for population dietary surveillance is discussed in a systematic review by Bandy et al., (2019), and these data are becoming increasingly utilised for the design and monitoring of policies which act upon the retail food environment to alter food purchase behaviours (Pell et al., 2021; Griffith et al., 2021; Griffith and O'Connell, 2009). Barcode scanning has a notable advantage over self-reported surveys like the FFS, as it automates the data collection and coding of product nutritional composition facilitating more accurate capture of purchased nutrients and permitting scale-up to many more participants. Yet, with the reliance on participants to scan their purchases, the method is not fully objective and may be open to selective omission of less healthy items. Panel data has been found to systematically under-estimate the purchase of soft drinks and snack products which are often eaten on the go and do not make it back to the participants' home for scanning (Einav, 2008). Whether such omissions are intentional or not, it is possible that the conscious need for participants to record their purchases may lead to changes in their purchasing behaviours. What is more, costs associated with data access may be prohibitive to their use in research.

Electronic supermarket purchase records, particularly those linked to a customer loyalty card, are another emerging method for collecting food purchase data. Their utility in dietary research is the subject of this thesis. Supermarket transactions may be limited by their coverage of purchases from only one retailer, missing food purchased out of home or elsewhere, but their secondary and objective nature is a notable advantage. Dubbed by some as 'accidental' (Arribas-Bel, 2014), or perhaps more accurately *incidental*, commercial secondary big data (including loyalty card transaction records) are inherently passive in nature which increases their objectivity and limits the Hawthorne effect. Purchases are collected continually providing data with a long-term longitudinal follow up, and a fine spatial and temporal granularity. They also capture a large proportion of the population. Unpublished results from the LifeInfo survey (Morris et al., 2018) suggest that almost 70% of the population hold at least one supermarket loyalty card, with many people holding cards for more than one retailer. Furthermore, large numbers offer a good coverage of low income groups who are traditionally hard to reach (Clark et al., 2021; Jenneson et al., 2022). As a result, researchers are beginning to see the appeal of supermarket transaction records for dietary monitoring (Jenneson et al., 2021). Yet with just a small number of studies investigating this (Eyles et al., 2010; Vepsäläinen et al., 2021), the validity of supermarket transactions in relation to more established dietary assessment methods is thus far not well understood. While the costs of collecting supermarket transaction records is not well documented, they are considered to offer a cost-effective means of dietary monitoring (the majority of costs being borne by the supermarket who already collect the data for business purposes), though the need to develop trusted retailer-academic data sharing agreements remains a barrier to their use in research.

At the Leeds Institute for Data Analytics we have developed such a partnership with Sainsbury's Plc (LIDA, 2021), one of the big four supermarkets in the UK with a 15% market share (Statista, 2021). Supermarkets are major retailers and producers of food in high and middle-income countries. In the UK, an estimated 78% of weekly expenditure on food and non-alcoholic beverages is made in large supermarkets (Office for National Statistics., 2021). As such, large supermarket brands have a lot of influence on the food environment (stage 1 in the FED). In recognition of their power to shape the nation's food-related habits, and the relatively limited capacity for population-level change through education and individual willpower alone, supermarkets (and the food industry as a whole) have received significant policy attention in recent years in the UK. To begin with, voluntary action through the setting of standards and targets (such as reformulation as part of the Public Health Responsibility Deal (DOH, 2015), and voluntary Multiple Traffic Light Labelling (Department of Health., 2017) for front of pack nutrition information) were the main policy levers employed. But with limited success, the Government has moved to a strategy of legislating mandatory actions. These include the Soft Drinks Industry Levy (SDIL) (HMRC, 2018) and forthcoming restrictions on the price and location-based promotions of foods high in fat, salt and sugar (HFSS) (GOV.UK, 2020b), with the possibility of further legislation on the horizon, such as a possible levy on added salt and sugar as proposed by the National Food Strategy (Griffith et al., 2021; National Food Strategy., 2021).

This spotlight on health, and more recently sustainability, has driven some retailers to take a more visible interest in their Corporate Social Responsibility (CSR). Industry-wide commitments to healthy and sustainable diets signify a shift in the narrative within the food industry (IGD, 2021a; The Consumer Goods Forum., 2020; The Consumer Goods Forum., 2021), catalysed by recent food system shocks experienced as a result of the COVID-19 global pandemic (Baty, 2020). Examples of partnership activities include the Peas Please pledge to increase the vegetable content of retail dishes (Food Foundation., 2020a; Food Foundation., 2020b), and a trial by Sainsbury's aimed at increasing purchases of fruits and vegetables (IGD, 2021b). The apparent willingness for industry-academic-policy partnerships and the use of data for pubic good (Baty, 2020) in today’s climate could be a real opportunity to advance the use of transaction records for population dietary research. Of course, like all businesses, supermarkets have competing commercial priorities which conflict with health and sustainability (National Food Strategy., 2021). Researchers working in partnership with the food industry should therefore be cautious of companies’ motives and power in influencing food research and policy (Nestle, 2003). That said, if a careful balance can be struck, the breadth and depth of data collected by supermarkets (about their products, customers, stores, sales etc) present a unique opportunity to learn about and shape the food system for the better.

With potential for supermarket transaction records to contribute to food system digitalisation (Baty, 2020), dietary surveillance, and the design, practical support and evaluation of food policy (Jenneson et al., 2021), there is much debate yet to be had around if and how retailers and academics can best collaborate to utilise transaction data for public good. While this is undoubtedly an interesting and necessary conversation, this thesis is concerned more with the methodological practicalities of applying transaction records in dietary research, than the philosophical debate around corporate power dynamics and alike. Nevertheless, it is important to consider the ethical issues in relation to data security, consent, and anonymity for working with supermarket transactions (a form of secondary big data). The specifics of how these ethical considerations impact the study protocols employed in this research are therefore discussed in Chapter 2, Data Preparation and Methods.

1.5 Thesis aim

This thesis aims to understand and evaluate the contribution that supermarket loyalty card transaction records can make to population dietary monitoring, both on a national and a small-area scale.

1.5.1 Objectives

The aim of this thesis is met through the following 9 objectives. Each objective is addressed by one of the three papers presented in the thesis.

Paper 1 – *A systematic review of automated electronic supermarket sales data for population dietary surveillance.* (Chapter 3)

1) Systematically review the literature on the use of electronically captured transaction data for dietary research.

2) Critically evaluate the strengths and weaknesses in their application to population dietary research.
3) Identify gaps in the evidence base, proposing areas for future novel research.

Paper 2 - *Exploring the geographic variation in fruit and vegetable purchasing behaviour using supermarket transaction data.* (Chapter 4)

4) Identify and collect a novel sample of detailed complete product-level purchase data spanning one year, from around 50,000 loyalty card customers of a leading UK supermarket.
5) Quantify average fruit and vegetable portions purchased at the household-level by matching retailer product categories to established food categories used in the Living Costs and Food Survey.
6) For the first time, apply regression analysis to supermarket transaction records linked to open-source population-level characteristics, to analyse geographical patterns in actual diet-related behaviours at the small-area level

Paper 3 - *A validation study: Supermarket Transaction Records In Dietary Evaluation (STRIDE).* (Chapter 5)

7) To design and recruit a sample of supermarket loyalty card holders to participate in a validation study in partnership with Sainsbury's, a large UK-based supermarket retailer, to assess agreement between dietary estimates from transaction records and self-reported intake.
8) To develop a bespoke nutrient composition database, using back of pack product data and national UK food tables, to link with transaction records for the calculation of absolute and energy-adjusted daily purchase estimates at the household and individual-level for energy, key macronutrients and sodium.
9) To quantify the statistical agreement (and limits to agreement) between dietary measures from loyalty card transactions and dietary measures from an online Food Frequency Questionnaire, assessing how agreement varies by customer subgroup.

1.6 Background chapter summary

In this chapter, I have described the challenges associated with population-level dietary monitoring and set out a valid place for supermarket transaction data to contribute to dietary assessment with the hope of overcoming some of these challenges. I have also introduced the aim and nine associated objectives of this thesis the success of which will be revisited in the final discussion chapter (Chapter 6). Next, in Chapter 2, I will introduce the data and methods used throughout the thesis, before addressing each objective listed through three research articles (Chapters 3 – 5), followed by a final discussion chapter reflecting on the overall aim of understanding the contribution that supermarket loyalty card transaction records can make to population dietary monitoring by assessing their validity.

Chapter 2
Data Preparation and Methods

This chapter supplements the methods described in the papers over the next three chapters, providing a more in-depth overview of the methods and data preparation. As the data preparation and methods undertaken to conduct the first paper (a systematic review of supermarket automated electronic sales data for population dietary surveillance) are described elsewhere in the systematic review protocol (Appendix A.1) and in the paper presented in Chapter 3, they are not included here. Instead, this chapter focuses on the preparation of secondary retail data used in both papers 2 and 3 (Chapters 4 and 5 respectively) and the design of the STRIDE study. As supermarket transaction records are not intended for use in research, the efforts and preparation required to repurpose them should not be underestimated. By describing the data preparation procedures here in detail, it is hoped that this thesis will contribute to the development of an established protocol for the routine and continued use of transaction records for population dietary assessment.

2.1 Retail transaction data (2016 sample)

The second paper contributing to this thesis is entitled *“Exploring the geographic variation in fruit and vegetable purchasing behaviour using supermarket transaction data”* (Jenneson et al., 2022), and can be found in Chapter 4. This investigation utilised secondary retailer data collected during the 2016 calendar year, for customers with loyalty cards registered in the Yorkshire and Humber region of England (henceforth referred to as the 2016 sample). Here I describe the large and novel, secondary data which made up the 2016 sample, before going on to describe the data which contributed to the STRIDE study in section 2.2.

The 2016 sample represents food and beverage (including alcoholic beverages) purchases for customers residing in the Yorkshire and Humber region in the north of England. As the largest English county, the study region was chosen to represent a broad range of geographic settings and customer demographics. The county is made up of vibrant multicultural cities in West and South Yorkshire which contain both prosperous and deprived communities in close proximity. The north and east of the region are more rural and contain affluent pockets as well as isolated village communities and an older age demographic, particularly on the east coast. Additionally, Yorkshire and Humber is home to the University of Leeds, where this study was conducted, making insights from the research highly relevant to the local community of which the University is part.

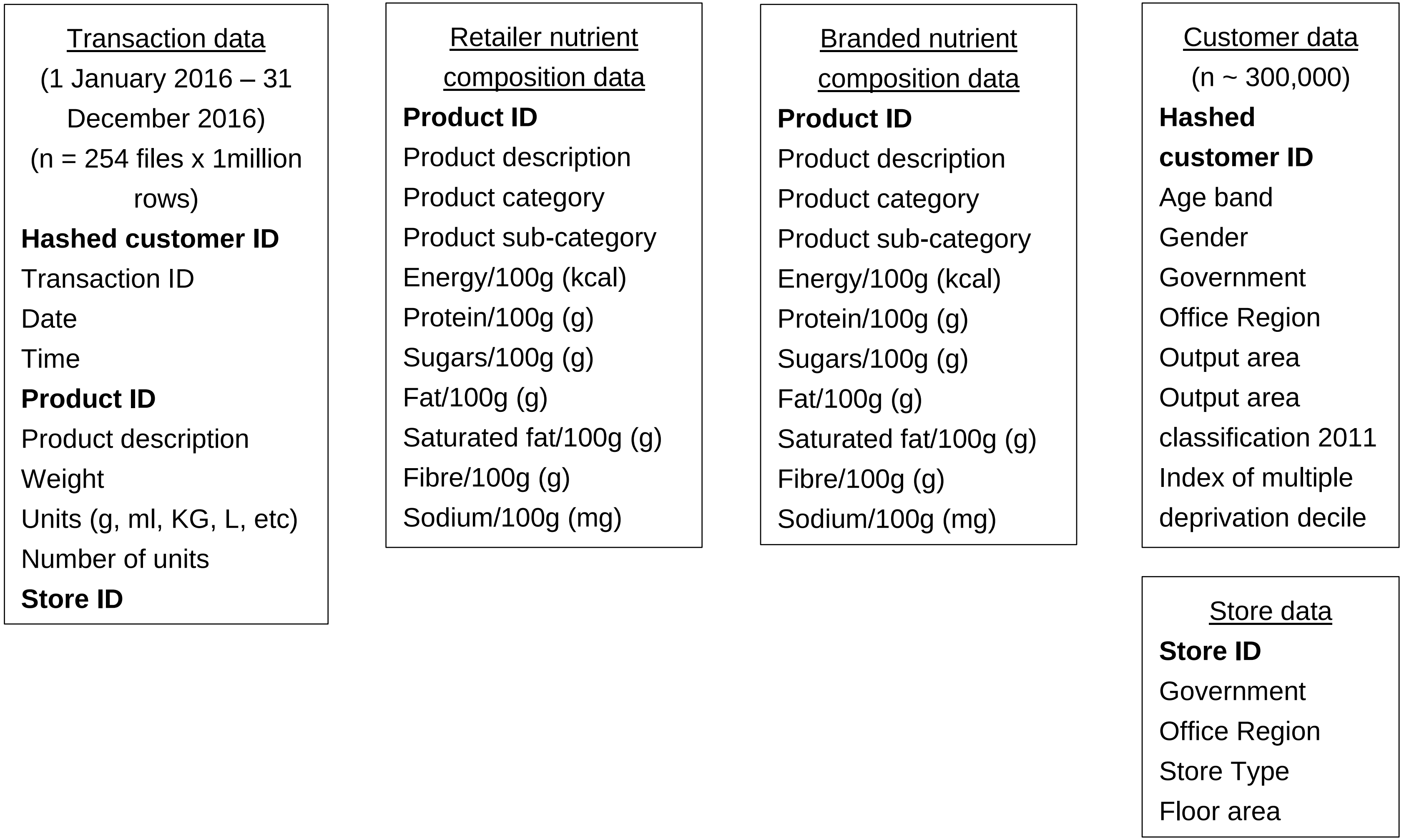
Customers in the sample were selected to represent “primary” shoppers; households who do the majority of their shopping with the retailer. Thus their purchase data is likely to represent their diet well and excludes customers who shop rarely with the retailer or only from a very limited range of categories (for example, customers who buy a meal deal only). Customer sampling criteria was therefore designed to set a minimum threshold for annual shopping frequency and breadth of categories, as described by Clark et al., (2021). It was considered feasible, particularly for those living in rural communities and those shopping online, that customers may do a ‘main’ shop once per month. With this in mind, it is also reasonable to allow monthly shoppers the flexibility to skip a couple of months, accounting for holidays away from home (e.g. at Christmas or during the summer). For this reason, a minimum threshold of 10 shopping trips per year, which each contained items from at least seven out of 15 food groups (or purchasing a ready meal and from three other categories), was set for inclusion. The 15 categories were based on the Living Costs and Food Survey (LCFS) categories (Office for National Statistics., 2017): Carbohydrate Products; Meat and Fish; Dairy; Fats; Fruit; Salad/Vegetables; Potatoes; Sweets; Other; Non-alcoholic drinks; Alcoholic drinks, plus an addition four categories derived by the research team: Ready Meals; Baby Food; Cakes and Biscuits; Crisps and Nuts (Clark et al., 2021).

The 2016 data sample was permitted for use by a suite of projects exploring customers food purchase behaviours, covered by approval from the University of Leeds ethical review board (reference number AREA – 18-050) together with a Data Licence Agreement (DLA) between the University of Leeds and Sainsbury’s Plc. An example of another project governed by the agreement is an investigation of dietary patterns using unsupervised machine learning (Clark et al., 2021). The DLA enables the use of the 2016 data sample by specified researchers at the Leeds Institute for Data Analytics (LIDA) for agreed projects. The legal basis for processing the data is expressed under the terms of the loyalty card sign-up agreement which Sainsbury’s customers choose to enter into when they register for a loyalty card. This loyalty card agreement states that Sainsbury’s may share anonymised customer data with third party organisations for research purposes. As such, customers are not required to give explicit consent for their anonymised data to contribute to this research.

A summary of the data files provided by the retailer is presented in Figure 2.1, where each box represents a file (or files) containing the data fields listed. Data fields highlighted in bold type indicate unique data linkage keys and thus represent the relations between files. Prior to sharing with the research team, identifiable customer information and all non-food and beverage transactions were removed. The customer loyalty card number was also replaced with a pseudonym hashed customer ID. As it was not possible to contact customers, no further information could be collected via surveys.

The transaction data consisted of 254 files each containing up to 1 million rows of data (245,479,086 rows in total), where each row represents a transaction for a given item by a customer on a particular shopping trip. Product nutrient composition data contained back of pack nutrient values given per 100g (or 100ml) of product. Separate files were provided for Sainsbury’s own brand products (pulled from their internal product database) and branded products (provided by Brandbank, a digital product content provider commonly used by retailers to build their ecommerce sites, under a separate licence agreement between Brandbank and the University of Leeds). As a result, the files had different formats and variable names, and contained different product category and sub-category structures. For consistency, existing categories were mapped to a new categorisation scheme based on the LCFS categories, as detailed in the paper in Chapter 4.

Data was provided for 326,087 loyalty card customers in the Yorkshire and Humber region. For the purposes of the spatial investigation presented in paper 2 (Chapter 4), this was sampled to 50,939 customers in the Leeds city region only. The sampling strategy is described in more detail in the paper (Jenneson et al., 2022) (Chapter 4), but briefly it was designed to include ‘primary’ shoppers who did the majority of their food shopping with Sainsbury’s, based on purchase frequency and coverage of product categories. Focusing the sample on a smaller geographic area enabled neighbourhood spatial patterns to be observed across a single diverse city.



Transaction data
(1 January 2016 – 31 December 2016)
(n = 254 files x 1million rows)
Hashed customer ID
Transaction ID
Date
Time
Product ID
Product description
Weight
Units (g, ml, KG, L, etc)
Number of units
Store ID

Retailer nutrient composition data
Product ID
Product description
Product category
Product sub-category
Energy/100g (kcal)
Protein/100g (g)
Sugars/100g (g)
Fat/100g (g)
Saturated fat/100g (g)
Fibre/100g (g)
Sodium/100g (mg)

Branded nutrient composition data
Product ID
Product description
Product category
Product sub-category
Energy/100g (kcal)
Protein/100g (g)
Sugars/100g (g)
Fat/100g (g)
Saturated fat/100g (g)
Fibre/100g (g)
Sodium/100g (mg)

Customer data
(n ~ 300,000)
Hashed customer ID
Age band
Gender
Government Office Region
Output area
Output area classification 2011
Index of multiple deprivation decile

Store data
Store ID
Government Office Region
Store Type
Floor area

Figure 2.1 Secondary retail data relations diagram (2016 sample).
Each box represents a file (or files). Relations between files are shown in bold type.

Each customer ID in the dataset represents a loyalty card. Customer demographic data therefore represents the person to which the loyalty card is registered. We assume this person to be the primary shopper for the household, though shopping may be carried out by other household members. For example, it is possible to obtain multiple copies of the same loyalty card for other household members to use. Equally, it is possible that members of the same household may hold separate loyalty cards. While the retailer tries to identify and link loyalty cards which are registered at the same address, it is possible that some households may be represented by more than one loyalty card. Further limitations of loyalty card data are that we do not know the size and composition of the household, and customers may forget or choose not to scan their loyalty card, so some transactions may not be captured. Additional demographic data was available at the neighbourhood level by linking national statistics to customers based on their output area of residence (ONS, 2017), a small area census geography made up of around 120 households. The Index of Multiple Deprivation decile (GOV.UK, 2015) and Output Area Classification (Gale et al., 2016) were used to give a sense of the characteristics of the neighbourhood in which customers live. However, we cannot know whether neighbourhood characteristics are wholly representative of the customers in our sample.

To ensure the security of customer information and commercially sensitive data, customer names were not shared, and loyalty card IDs were replaced with unique customer pseudonyms (hashed customer ID) by the retailer before sharing with the research team. Data was shared via a secure file transfer platform and stored in a secure cloud-based research data environment maintained by the Data Analytics Team (DAT) in LIDA, which has ISO27001 and NHS DSPt accreditation, along with an independent 3rd party assessment by Sainsbury's. The secure environment was not connected to the internet and was accessible only by approved researchers. Further detail explaining how the data security was built into research procedures can be found in section 2.2.1.1 (Data Governance).

2.2 The STRIDE Study (Supermarket Transaction Records In Dietary Evaluation)

The STRIDE study was conducted to assess the validity of supermarket transactions as a population dietary assessment tool, against a more established measure of self-reported intake using an online Food Frequency

Questionnaire (FFQ). Addressing the thesis' major aim (and meeting objectives 7 - 9 set out at the end of Chapter 1), the results of the STRIDE study can be found in the third paper of this thesis (Chapter 5). With few assessments of the validity of supermarket loyalty card transaction records as a dietary assessment measure, the STRIDE study makes a novel contribution to this emerging field. In the subsequent sections of this chapter, I will provide further detail of the design, recruitment, data and methods used for the STRIDE study, to complement the paper. The retailer data files used in the STRIDE study are the same in structure and linkage as those used in the 2016 retailer sample (described in section 2.1). However, with the active recruitment of a study sample and inclusion of primary data collection, the STRIDE study took additional considerations into account, as described in the remainder of this chapter.

2.2.1 Designing the STRIDE study

Supermarket loyalty card transaction records represent a novel secondary big data source which is not readily available to researchers. The commercial sensitivity of the data, as well as the need for retailers to operate in respect of due diligence to the protection of customer's information, prohibits open access to these forms of data. While the secondary use of data in the 2016 Yorkshire and Humber sample (described earlier in this chapter in section 2.1) was permitted under the loyalty card sign-up agreement, the STRIDE study involved primary data collection and linkage of primary and secondary data sources. It was therefore necessary to gain informed consent from STRIDE study participants. Access to this data was made possible due to a trusted relationship and data governance procedures between the University of Leeds and Sainsbury's. The remainder of this chapter gives an overview of the design and processes which contributed to the STRIDE study, in greater depth than that which is found in the paper in Chapter 5.

2.2.1.1 Data Governance

Prior to the start of the STRIDE study project, the research team and retailer entered into a contract governed by a studentship agreement, data agreement and non-disclosure agreement designed to protect customers and the commercial sensitivity of retailer data. In addition, the project underwent a series of due diligence checks. The protocol for the STRIDE study (Jenneson et al., 2020) can be found online via the Open Science Framework, and the Data Management Plan (DMP) is in Appendix B.1 at the end of this thesis. The protocol and DMP received approval from; the University of Leeds ethical review board (reference number AREA – 18-174) (Appendix B.2); Sainsbury's internal data clinic (which ensures data shared with third party organisations is to be used safely and in accordance with the company's values); and the LIDA Data Analytics Team (who maintain the secure data facility used for the study). To ensure data security was maintained, the project was designed to meet the Five Safes model (Desai, 2016) as described below.

1. **Safe data**: Data was pseudonymised using a unique study ID for each participant. Identifiable information (email address) was held separately from FFQ and transaction data to prevent linkage and identification of customers.

2. **Safe projects**: The project was approved by the University of Leeds ethical review board as well as the data owner via their in-house ethical review procedures. This comprised a due diligence review to minimise risk and maximise public good, including ensuring data collection tools met required procurement standards for evidence of regular penetration testing.

3. **Safe people**: Data could only be accessed by named members of the researcher team who were trained in safe data access protocols. Training was refreshed annually.

4. **Safe settings**: Data was held in LIDA's SecureLab environment LASER. This permitted offline data access in a secure environment for approved researchers. Retailer data was transferred to LIDA via a secure file transfer platform (Biscom) which offers an air-locked temporary landing space. Inputs were reviewed by LIDA's Data Analytics Team against the study protocol to ensure they met the project's agreed remit and did not pose a risk for statistical disclosure.

5. **Safe outputs**: Outputs from the secure data environment were screened and approved by the Data Analytics Team to ensure they are non-disclosive and in line with the project remit. Outputs were additionally screened for commercial sensitivity by the retailer team.

2.2.2 STRIDE study recruitment

This section compliments the STRIDE paper by describing in more detail the participant recruitment journey. Learnings from the pilot phase which led to changes to the protocol that influenced the final study design, are also summarised.

Similar to the 2016 retail data sample (described in section 2.1), participants of the STRIDE study were selected to be 'primary' shoppers, for whom purchases at Sainsbury's were likely to represent the majority of their food shopping. An eligible sample was selected based on the frequency (at least 10 times per year) and breadth (from at least seven out of 15 pre-defined categories derived from the Living Costs and Food Survey categories (Office for National Statistics., 2017)) of their purchases in the 2019 calendar year. The frequency criterion was designed to capture customers who may do a main shop once per month (expected to be more common among customers in rural communities who live further from a supermarket, and those who do online shopping), allowing for a couple of months off, for example where they may go away for the holidays. The breadth criterion aims to include customers for whom purchases at Sainsbury's capture the majority of their diet, while allowing for dietary exclusions (e.g. people who follow a vegan diet will not shop in dairy, fish or meat categories).

As described in the study protocol (Jenneson et al., 2020), the STRIDE study was designed to recruit a total of 2,250 people across all cohorts (including the pilot phase) from the Yorkshire and Humber region of England. Similar studies collecting purchase data and multiple 24-hour dietary recalls had achieved very high completion rates (between 75 – 96%) (Ransley et al., 2001; Eyles et al., 2010; Wark et al., 2018), upon which an anticipated completion rate of 80% for the STRIDE study was based. This was expected to return around 1,800 participants with complete FFQs and transaction records. This number was chosen to give sufficient power for subgroup analyses; it is recommended that around 200 datapoints be used for the Bland and Altman test for statistical agreement (Bland, 1986)). Based on the 2011 census, 88.8% of people in the Yorkshire and Humber region reported their ethnicity as 'White', with the remaining 11.2% of 'Non-white' ethnicity (GOV.UK, 2018)). Assuming the study sample is representative of the Yorkshire and Humber as a whole, I estimated around 1,800 people would therefore be required to enable subgroup analysis by ethnicity ((200/11.2)*100 = 1,786).

From conversations with retail colleagues a maximum recruitment rate of 5% was anticipated based on industry expectations around customer engagement with market research communications. The anticipated recruitment and completion figures for each cohort are given in Table 2.1. The pilot phase was completed in May 2020 and provided an opportunity to assess the study protocol and learn from the experience in order to make adjustments to the full study protocol.

Table 2.1 STRIDE expected recruitment figures (pre-pilot)

Study period	**Pilot**	**Cohort 1**	**Cohort 2**	**Cohort 3**	**Cohort 4**	**TOTAL**
Invited	5,000	10,000	10,000	10,000	10,000	45,000
Expected sign-up (5%)	250	500	500	500	500	2,250
Expected completion (80%)	200	400	400	400	400	1,800

The participant journey for the pilot phase is summarised in Figure 2.2. Participants were invited via email from the partner retailer to participate in the STRIDE study. The email was personalised for each customer and included a link to the study website where customers could review participant information (prepared in accordance with the University's ethical guidance for informed consent). Participant information is shown in full in Appendix B.3. From the study website, participants could then click another link to visit the online sign-up and baseline survey form. Wording for the informed consent form is provided in full in Appendix B.4. Around 1-2 weeks after completion of the consent form, I extracted the email addresses of participants from the consent form and used these to share a personalised link to the online FFQ.

Actual recruitment figures from the pilot phase were reviewed alongside web analytics tools for the study website and the online survey (consent and baseline). The pilot phase yielded a sign-up rate of 1.6% (n = 80) and a completion rate of 49% (n = 39 with complete FFQs). Just 1.25% of those who signed up actively withdrew from the study, while the remaining 40 people who failed to complete were lost to follow up. This gave an overall completion rate of 0.78% (39/5,000), much lower than the expected 4% (200/5,000).

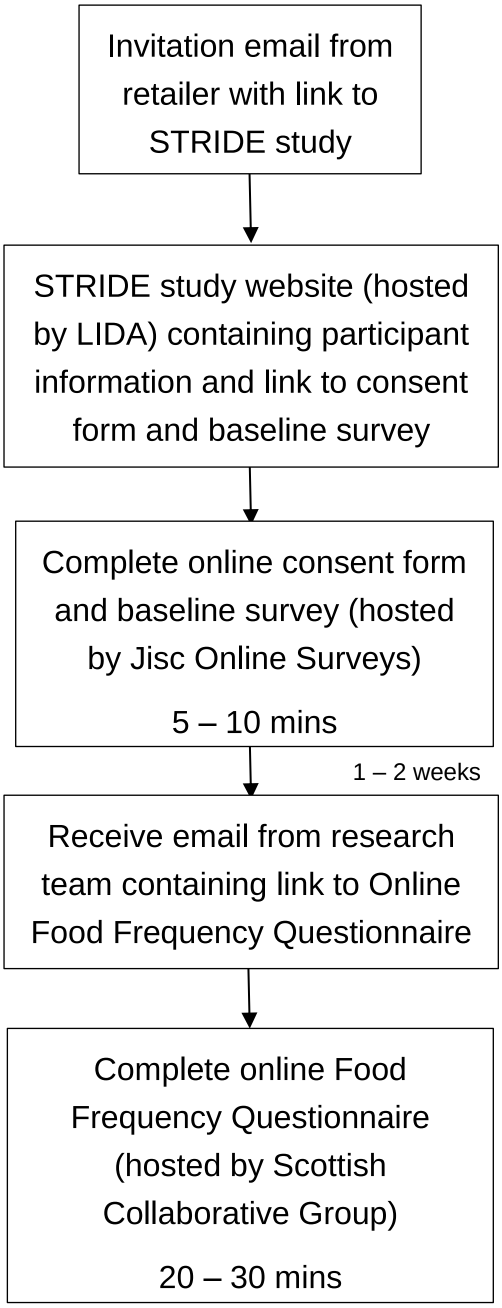
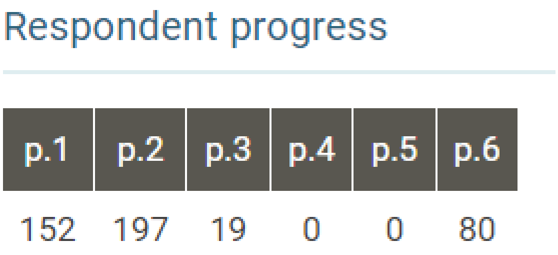

Invitation email from retailer with link to STRIDE study
STRIDE study website (hosted by LIDA) containing participant information and link to consent form and baseline survey
Complete online consent form and baseline survey (hosted by Jisc Online Surveys)
5 – 10 mins
1 – 2 weeks
Receive email from research team containing link to Online Food Frequency Questionnaire
Complete online Food Frequency Questionnaire (hosted by Scottish Collaborative Group)
20 – 30 mins

Figure 2.2 Participant journey flow diagram for STRIDE pilot study

Web analytics revealed that the STRIDE study web page received 635 unique page views during the pilot phase, indicating that 12.7% of customers who received the invitation email were interested enough to visit the study page. Average length of time on the page was just 15 seconds, indicating that people quickly left the page without much engagement with its content. A screenshot from analytics outputs of the baseline survey from the pilot phase is found in Figure 2.3, which shows that in total 448 people visited the survey. While 80 people left the survey on the final page (having completed the sign-up), the majority of people left on the first two pages. These findings suggest that while there was a reasonable amount of curiosity about the STRIDE study our ability to convert this to sign-up was poor.



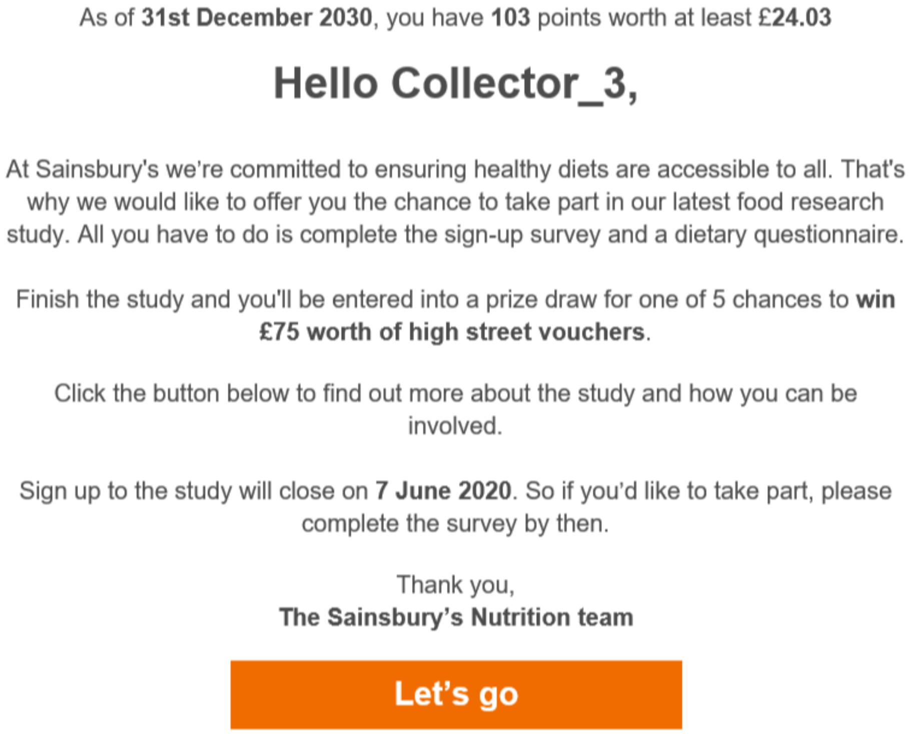
Respondent progress

p.1	p.2	p.3	p.4	p.5	p.6
152	197	19	0	0	80

Figure 2.3 STRIDE pilot baseline survey analytics

This is a figure showing the number of people leaving the baseline survey on each of the six pages. People leaving before page 6 did not complete the sign-up process.

Following the pilot phase, a number of changes were made to improve the participant experience of the sign-up process. The participant journey was simplified by changing the destination of the sign-up button in the invitation email to the consent survey rather than the study webpage. The sign-up button was also made more prominent in the sign-up email. A screenshot of a test version of the sign-up email is included in Figure 2.4, which also included terms of the prize draw and links to privacy policies underneath (Figure 2.5).



As of **31st December 2030**, you have **103** points worth at least £**24.03**

Hello Collector_3,

At Sainsbury's we're committed to ensuring healthy diets are accessible to all. That's why we would like to offer you the chance to take part in our latest food research study. All you have to do is complete the sign-up survey and a dietary questionnaire.

Finish the study and you'll be entered into a prize draw for one of 5 chances to **win £75 worth of high street vouchers**.

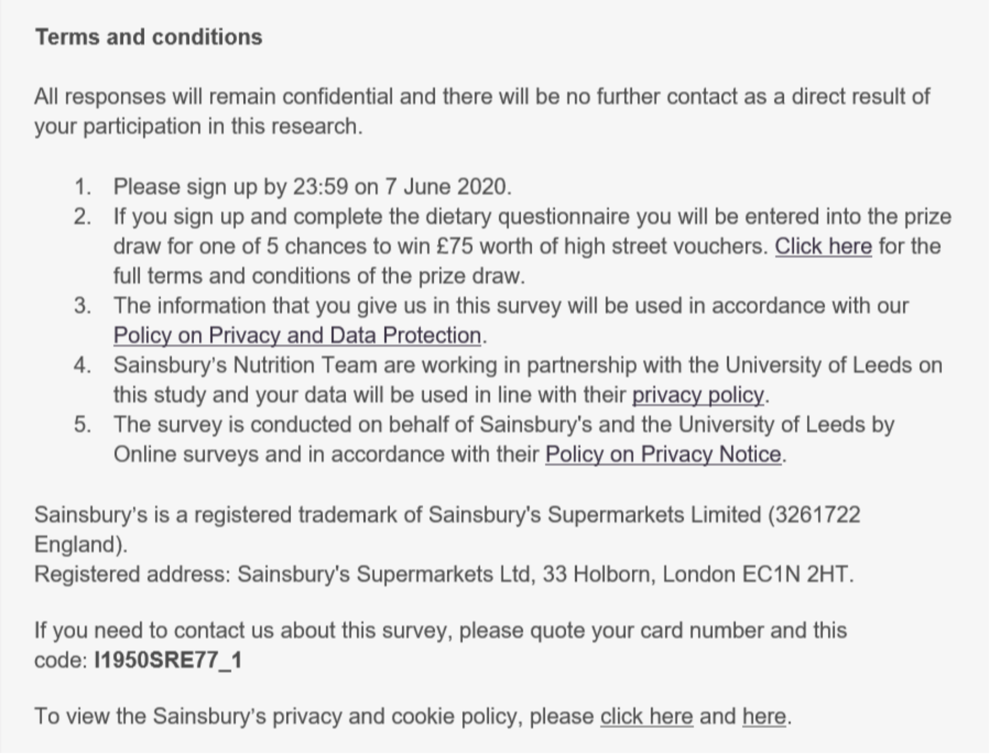
Click the button below to find out more about the study and how you can be involved.

Sign up to the study will close on **7 June 2020**. So if you'd like to take part, please complete the survey by then.

Thank you,
The Sainsbury's Nutrition team

Figure 2.4 STRIDE invitation test email

This is a figure showing the main body of the email sent to customers by Sainsbury's inviting them to take part in the STRIDE study. Emails were personalised with the customer's name and the closing date was updated for each cohort.



Terms and conditions

All responses will remain confidential and there will be no further contact as a direct result of your participation in this research.

1. Please sign up by 23:59 on 7 June 2020.
2. If you sign up and complete the dietary questionnaire you will be entered into the prize draw for one of 5 chances to win £75 worth of high street vouchers. Click here for the full terms and conditions of the prize draw.
3. The information that you give us in this survey will be used in accordance with our Policy on Privacy and Data Protection.
4. Sainsbury's Nutrition Team are working in partnership with the University of Leeds on this study and your data will be used in line with their privacy policy.
5. The survey is conducted on behalf of Sainsbury's and the University of Leeds by Online surveys and in accordance with their Policy on Privacy Notice.

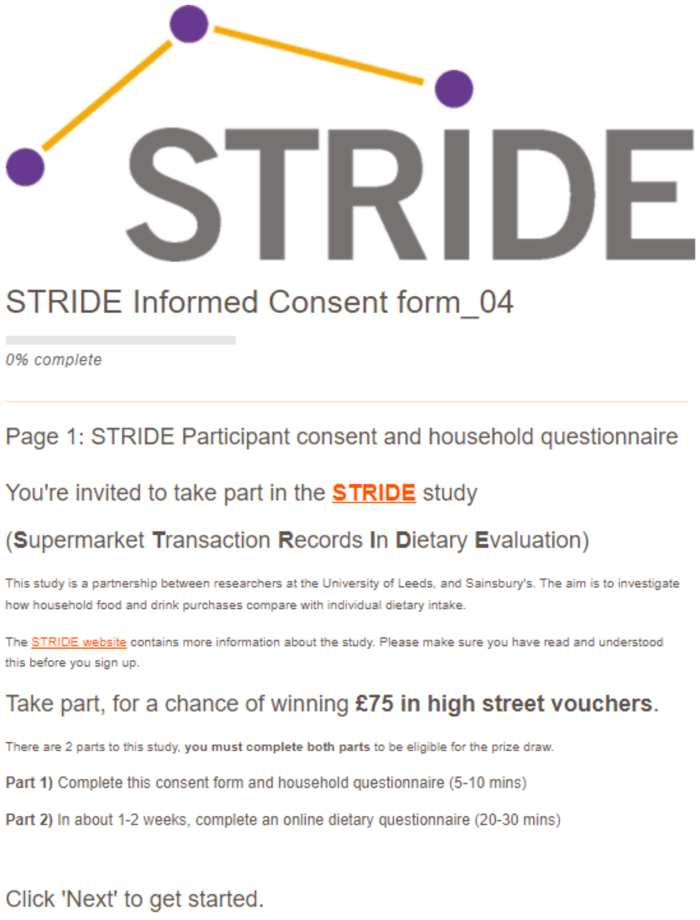
Sainsbury's is a registered trademark of Sainsbury's Supermarkets Limited (3261722 England).
Registered address: Sainsbury's Supermarkets Ltd, 33 Holborn, London EC1N 2HT.

If you need to contact us about this survey, please quote your card number and this code: **I1950SRE77_1**

To view the Sainsbury's privacy and cookie policy, please click here and here.

Figure 2.5 STRIDE invitation email terms and links to privacy notices

To attract more sign-ups, wording on the survey homepage (Figure 2.6) was updated to be more concise and inviting. The incentive and links to the study webpage were also made more prominent. Additionally, it was made clearer up front that participants would be signing up to a 2-step process which included the online FFQ. Wording was also changed on the study webpage to be more concise and inviting (Figure 2.7) whilst incorporating buttons to make links to the survey and reference to the incentive more prominent. The layout of the webpage was changed to include a list of expandable headings outlining 'more information' (Figure 2.8), to enable visitors to quickly see what information was provided and navigate to areas they wanted to read, without having to scroll down the whole page.



STRIDE

STRIDE Informed Consent form_04

0% complete

Page 1: STRIDE Participant consent and household questionnaire

You're invited to take part in the **STRIDE** study

(**S**upermarket **T**ransaction **R**ecords **I**n **D**ietary **E**valuation)

This study is a partnership between researchers at the University of Leeds, and Sainsbury's. The aim is to investigate how household food and drink purchases compare with individual dietary intake.

The STRIDE website contains more information about the study. Please make sure you have read and understood this before you sign up.

Take part, for a chance of winning **£75 in high street vouchers**.

There are 2 parts to this study, **you must complete both parts** to be eligible for the prize draw.

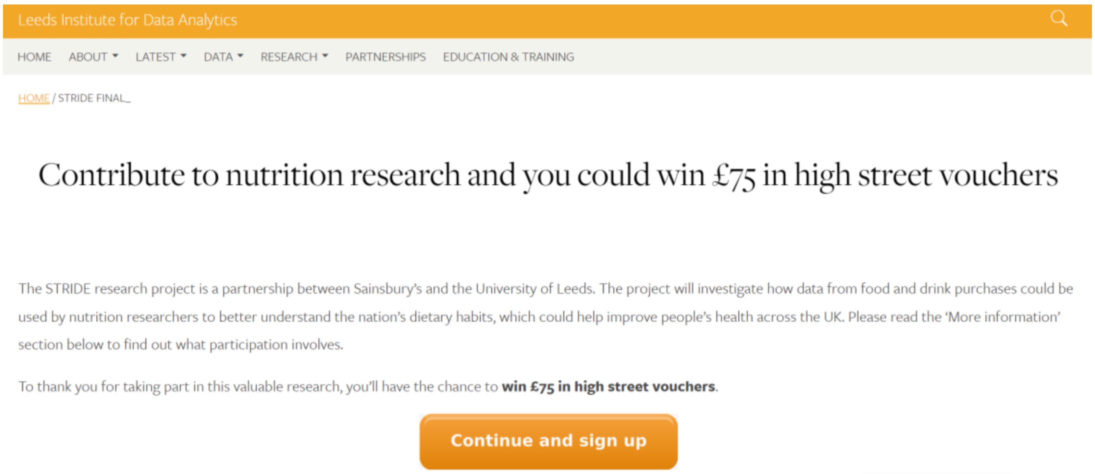
Part 1) Complete this consent form and household questionnaire (5-10 mins)

Part 2) In about 1-2 weeks, complete an online dietary questionnaire (20-30 mins)

Click 'Next' to get started.

Figure 2.6 STRIDE baseline survey and consent form homepage

This is a figure showing the landing page for the STRIDE baseline survey hosted by Jisc Online Surveys. It contains links to participant information on the study website.



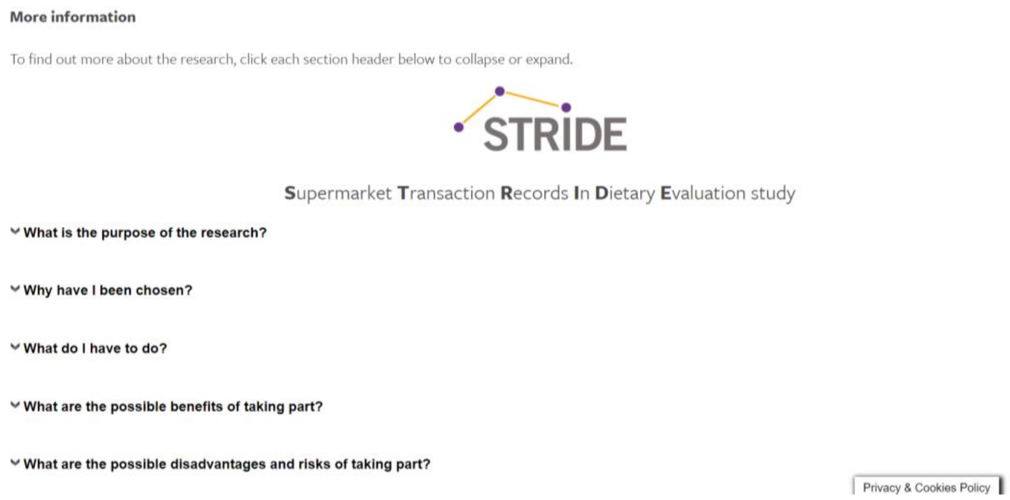
Contribute to nutrition research and you could win £75 in high street vouchers

The STRIDE research project is a partnership between Sainsbury's and the University of Leeds. The project will investigate how data from food and drink purchases could be used by nutrition researchers to better understand the nation's dietary habits, which could help improve people's health across the UK. Please read the 'More information' section below to find out what participation involves.

To thank you for taking part in this valuable research, you'll have the chance to **win £75 in high street vouchers**.

Figure 2.7 Screenshot of STRIDE study webpage containing participant information

This is a figure showing the introductory section on the STRIDE study webpage. It introduces the study and includes a link to the baseline and consent survey.



More information

To find out more about the research, click each section header below to collapse or expand.

STRIDE

Supermarket **T**ransaction **R**ecords **I**n **D**ietary **E**valuation study

- **What is the purpose of the research?**
- **Why have I been chosen?**
- **What do I have to do?**
- **What are the possible benefits of taking part?**
- **What are the possible disadvantages and risks of taking part?**

Figure 2.8 Screenshot of STRIDE study webpage – More information headings

This is a figure showing the expandable headings on the STRIDE study webpage. Participants could click each heading to read the participant information (given in full in Appendix B.3).

In addition to the outlined changes to the user journey it was discovered that the rate of linkage of loyalty card IDs provided at sign-up to customer transaction records from the pilot phase was 91%. Unlinked customers were thought to be due to error when participants entered their long card number from the front of their loyalty card. Furthermore, due to differences in internal customer IDs used across the business, only half of customers could be matched to customer information on the loyalty card database. To improve match rates, the unique link to the baseline survey in the invitation email was embedded with a second unique hashed ID for each customer. This auto-populated a field in the baseline questionnaire with the hashed ID, in addition to the manually entered loyalty card number. With two unique identifiers available from which to find the customer record, successful identification increased to an average of 97% across the remaining four cohorts.

Finally, the expected sign-up rate was revised to around 1 – 1.5% and the completion rate revised to around 50%. Based on this, the number of recruitment emails required to be sent increased four-fold from 10,000 to 40,000 per cohort (Table 2.2). As a result, it was necessary to lower the expected sample size to around 1,000 and expand the sampling frame to four regions in England (Yorkshire and the Humber, East Midlands, West Midlands and the South East) rather than just the Yorkshire and Humber region as originally planned.

Table 2.2 Revised expected sign-up and completion figures for STRIDE cohorts

Figures were revised based on sign-up and completion figures observed in the pilot phase. This included increasing mail-out size and reducing the expected sample size.

Study period	**Pilot**	**Cohort 1**	**Cohort 2**	**Cohort 3**	**Cohort 4**	**TOTAL**
Invited	5,000	40,000	40,000	40,000	40,000	165,000
Sign-up rate (1 – 1.5%)	80	500	500	500	500	2,080
Completion rate (50%)	39	250	250	250	250	1,039

2.2.3 STRIDE data collection

Participants of the STRIDE study were asked to complete a baseline questionnaire at the beginning of the study, followed by a FFQ around 1-2 weeks later. In addition, participants consented for their loyalty card transaction records to be retrospectively shared with the research team. Each of the three data collection methods is next described in turn.

2.2.3.1 Baseline questionnaire

The baseline questionnaire was combined with the consent form to create a smoother participant sign-up journey. The survey was hosted by Jisc Online Surveys, an approved survey provider for academic research to which the University of Leeds subscribes. The full list of baseline questions is provided in Appendix B.5. In addition to standard demographic questions, participants were asked about their household composition for the purpose of calculating individual-level purchase estimates (described in more detail later in section 2.2.5). Data preparation included the calculation of participant age at sign-up and the calculation of Body Mass Index (BMI) from self-reported height and weight. This required some data cleaning to ensure all data entries were in the correct units (metres and kilograms respectively).

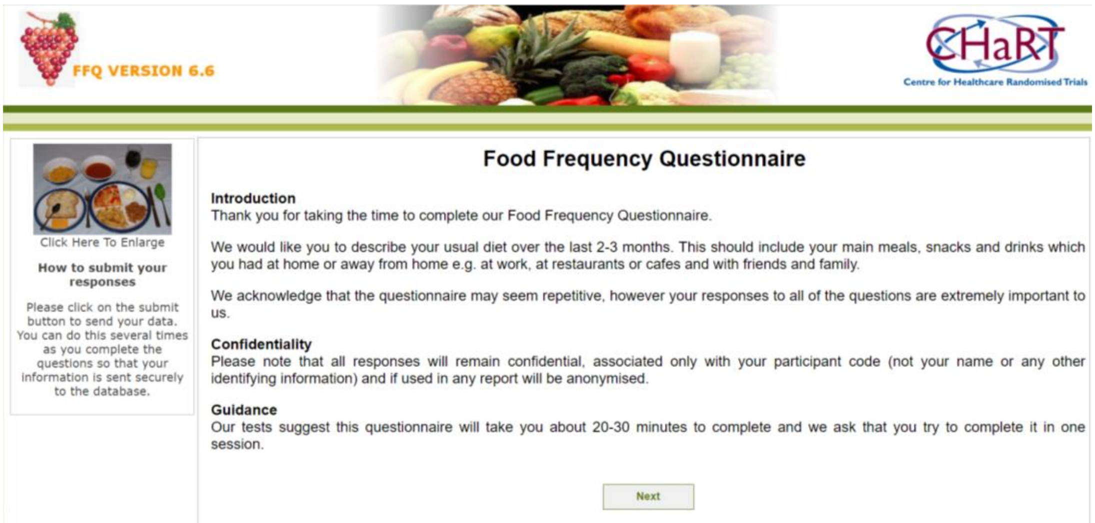
2.2.3.2 Dietary intake assessment

Nutritools., (2019) (a website summarising dietary assessment tools) was used to identify validated online dietary assessment tools for use in the STRIDE study. It was important that dietary assessment should be carried out online to maximise the possible sample size whilst minimising costs. Furthermore, the retailer advised that online communications are the dominant mode of communication for their loyalty card customers, and that they are used to completing online questionnaires. Shortlisted tools were validated for use in the UK adult population to monitor intake of energy and macronutrients. It was not considered important to capture micronutrient intake as these are not included on product nutrition labels (with the exception of sodium) for direct comparison with purchase data.

The tools considered for use were; myfood24 (Carter et al., 2015), INTAKE24 (Simpson et al., 2017), Oxford WebQ (Liu et al., 2011) and the Scottish Collaborative Group (SCG) FFQ (Masson et al., 2003). To meet the retailer’s procurement standards for questionnaire tools (typically applied in a market research context rather than for academic research) it was stipulated that the chosen tool must conduct regular penetration (PEN) testing of their platform. PEN testing is used to ensure the security of participant data on the platform against malicious threats and was therefore part of the due diligence requirement for duty of care towards retailer customer data. Enquiries were sent to the research teams responsible for each of the shortlisted tools to ascertain costs for use and evidence of PEN testing. At the time of design, just one tool, the SCG FFQ, could provide evidence of regular PEN testing and was therefore selected as the dietary assessment tool for the STRIDE study.

As the only online FFQ shortlisted, the SCG tool also has an advantage over the other tools that it captures ‘usual’ diet for the medium-term (2-3 months) without repeated completion, reducing participant burden. Participants were sent a unique link to complete the online FFQ around 1-2 weeks after completion of the consent form and baseline questionnaire. The FFQ is a 150-item semi-quantitative tool which asks respondents to estimate the amount of each food and how often they consume it (Masson et al., 2003). It has been validated against weighed food diaries (Masson et al., 2003; Jia et al., 2008), estimated food diaries (Mohd-Shukri et al., 2013; Hollis et al., 2017), and biomarkers (Heald et al., 2006), in UK adult and older adult populations.

A screenshot of the FFQ landing page is shown in Figure 2.9. This introduces what is expected of participants before taking them to a more detailed set of completion instructions. Page 1 of the FFQ is shown in Figure 2.10. Participants were required to complete all 9 pages of the FFQ, which are presented by food group. The amount and frequency of each food item was entered using dropdown boxes to state the number of measures per day and the number of days per week on which the item is consumed. If a participant did not consume the item they could select ‘R’ in the ‘Number of days per week’ column for ‘Rarely or never’.



How to submit your responses

Please click on the submit button to send your data. You can do this several times as you complete the questions so that your information is sent securely to the database.

Food Frequency Questionnaire

Introduction
Thank you for taking the time to complete our Food Frequency Questionnaire.

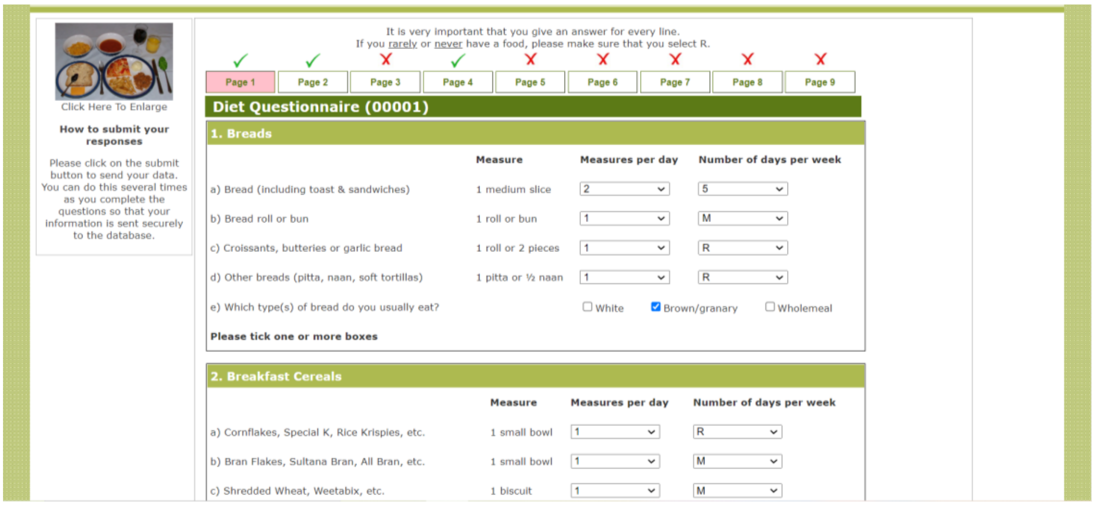
We would like you to describe your usual diet over the last 2-3 months. This should include your main meals, snacks and drinks which you had at home or away from home e.g. at work, at restaurants or cafes and with friends and family.

We acknowledge that the questionnaire may seem repetitive, however your responses to all of the questions are extremely important to us.

Confidentiality
Please note that all responses will remain confidential, associated only with your participant code (not your name or any other identifying information) and if used in any report will be anonymised.

Guidance
Our tests suggest this questionnaire will take you about 20-30 minutes to complete and we ask that you try to complete it in one session.

Figure 2.9 Screenshot of Scottish Collaborative Group online Food Frequency Questionnaire (SCG-FFQ) landing page



How to submit your responses

Please click on the submit button to send your data. You can do this several times as you complete the questions so that your information is sent securely to the database.

It is very important that you give an answer for every line.
If you rarely or never have a food, please make sure that you select R.

Diet Questionnaire (00001)

1. Breads

	Measure	Measures per day	Number of days per week
a) Bread (including toast & sandwiches)	1 medium slice	2	5
b) Bread roll or bun	1 roll or bun	1	M
c) Croissants, butteries or garlic bread	1 roll or 2 pieces	1	R
d) Other breads (pitta, naan, soft tortillas)	1 pitta or ½ naan	1	R
e) Which type(s) of bread do you usually eat?		□ White	☑ Brown/granary □ Wholemeal

Please tick one or more boxes

2. Breakfast Cereals

	Measure	Measures per day	Number of days per week
a) Cornflakes, Special K, Rice Krispies, etc.	1 small bowl	1	R
b) Bran Flakes, Sultana Bran, All Bran, etc.	1 small bowl	1	M
c) Shredded Wheat, Weetabix, etc.	1 biscuit	1	M

Figure 2.10 Screenshot of SCG-FFQ Page 1

The FFQ takes approximately 20 minutes to complete and does not need to be completed in one go (although this is recommended). Participants who did not complete the FFQ received up to two email reminders with a 1-2-week interval between. While it is possible to automate the sending of invitation and reminder emails within the SCG-FFQ tool, consent for the STRIDE study did not permit anyone outside the immediate research team at the University of Leeds (Mrs Victoria Jenneson and Dr Michelle Morris) to have access to identifiable participant information. Therefore, it was not possible to upload email addresses to the FFQ platform for automatic mailing. Instead, all emails were sent to the STRIDE study mailbox and forwarded to the correct recipient. This was a time-consuming process which required close care and attention. The STRIDE mailbox also required regular monitoring for participant support and notification of withdrawals. Participants who withdrew from the study were considered lost to follow up and were not contacted further. Unless participants explicitly requested for all their data to be withdrawn from the study, their baseline data was retained.

Once FFQ entries were submitted by participants they were each manually checked for completeness before sending to the SCG team for estimation of daily nutrient intakes. The SCG require that no more than 10 items on the FFQ be incomplete for analysis. Participants could use the free-text section at the end of the tool (question 20, ‘other foods’) to list any regularly consumed food or beverage items that they did not consider were captured by the FFQ. The form also asks for the quantity and frequency of ‘other foods’ to be entered manually. Other foods were manually checked and coded to the nearest item on the tool where applicable. For example, rosé wine was coded as white wine, oat milk was coded as soya milk, and gluten free bread was coded as standard bread. Alternatives were chosen where they were considered to be close to the stated ‘other food’ product in terms of their energy and macronutrient composition. Where no close alternative could be found on the form, the WinDiets Professional tool (2013 version) was used to find the CoFID 2008 (Composition of Foods Integrated Dataset) code (Public Health England., 2015) for the food in the UK National Composition of Food tables, which was reported along with the quantity to the SCG team.

Fats used in cooking and on bread were reported in question 17 of the tool. Participants were asked to state the two main types of fats used for cooking and the two main types used for bread (if any), along with the brand name if possible. A fat coding sheet (Appendix B.6) was provided to me by the SCG team and used to manually code each fat entry to a pre-defined list of 35 fat and oil types before entries were shared with the SCG team for analysis. A nutrient breakdown for each participant was provided by the SCG team, more information on its contents is provided in section 2.2.5.1.

2.2.3.3 Food purchase data

Transaction records were provided by the retailer for each study cohort. Participants' transactions were identified via their loyalty card ID (either the manually entered long card number, or the auto-populated hashed customer ID). Of the 1,768 who signed up across all cohorts (after removal of withdrawn participants (n=20), transaction records were obtained for 1,588 (a 90% match-rate). Transactions could not be found for all customers. Given that customers were selected based on their 2019 purchases, transaction records for the baseline period should have been available for all customers at the least. Therefore, the loss of 10% is expected to be caused by errors with the customer ID, due to human error when typing in the loyalty card number, or accessing the baseline questionnaire via the link from the study website causing the hashed customer ID not to be auto-populated. Transactions for each participant were provided retrospectively for the 1-year baseline period (1 January 2019 – 31 December 2019) and for the 1-year STRIDE study period (1 June 2020 – 31 May 2021). In total, the transactions data for the whole study (baseline and study period for all four cohorts plus the pilot phase) contains 2,959,012 rows of data and 208,194 unique transactions. The 3-month period of transactions which correspond to the period covered by the FFQ for each cohort is referred to as the cohort period.

The transaction data files were provided in long format csv files. Each line in the data represents an item purchased (identified by the product description and unique identifiers) by a participant (identified by the STUDY_ID) in a given transaction (identified by the transaction ID). The item weight, quantity purchased, time and date of transaction, amount paid (GBP £), ID for the store, and whether the item was purchased online or in store were also provided. The transaction file did not contain any nutritional information and therefore linkage with a separate product nutrient composition file was required.

2.2.4 Development of a bespoke nutrient composition database

One of the benefits of electronically captured purchase records is the ability to record the exact product purchased, and the amount. Automated linkage with product-specific nutrient composition data in theory permits a highly accurate estimation of purchased nutrients. In comparison, most estimations of consumed nutrients are reliant on matching recorded foods to a more limited list of generic food items in national food composition tables. By harnessing product nutrient composition data from the back of pack (BOP) nutrition panel at the unique product level, estimates of purchased nutrients remain up to date to changes in product recipe and new product launches, and are sensitive to brand-specific differences in composition.

BOP nutrient composition data was provided by Sainsbury's for all products sold in 2019. This contained quantities for energy (kcal), fat (g), saturated fat (g), protein (g), sugar (g) and sodium (mg) stated per 100g or per 100ml of product for 25,405 products. Product nutrition information was matched to products in the transaction files via the stock-keeping unit (SKU) or European Article Number (EAN). This returned a match rate of 72%. 7,034 products (equivalent to 28% of products) did not have a match in the product nutrient composition file; these are referred to as 'unmatched' products. Of those products with a matched entry in the product nutrient composition file, 5,566 (22%) products had missing or zero entries in the energy (Kcal/100g) field, these are referred to as 'missing' products. Inspection of the data revealed that where energy values were missing, other nutrient values also tended to be missing, so missing or zero energy values were considered a marker of 'missing nutrition data'.

Products for which there was no nutrient data in the product nutrition file (unmatched and missing) were manually coded to generic composition values in the CoFID (PHE, 2020). After imputation of missing products, a match rate of 96% was achieved (109,809 of transactions were matched in this way). The remaining 4% of product matches were achieved through imputation of unmatched products, taking the total match rate to 100% of products. Products with unmatched or missing product nutrient composition data were typically unpackaged fresh produce (e.g. fruits and vegetables, in-store bakery products, or deli counter items), alcoholic beverages (which are not required by law to display nutritional information on product packaging (Department of Health., 2017)), or seasonal items such as Easter eggs or Christmas treats. The CoFID database (PHE, 2020) was manually searched by name for a close match to the product, and data was imputed in Excel. Selection of the closest matching product relied upon my nutritional expertise. Where different options existed in the CoFID database for unprepared or prepared food items, values for the unprepared item were taken. This reflects that preparation method for the food item is unknown (for example, a potato may be peeled and fried as chips or it may be baked with its skin on, among other preparation methods).

Additionally, the weight of produce is 'as sold' e.g. the weight of bananas sold is inclusive of their skin, therefore the nutritional value selected matched this. Finally, I take the assumption that manufacturers will state the product nutritional values as sold, unless cooking or reconstitution instructions are stated on the pack (for example, for instant noodles where the manufacturer states how much water to add).

2.2.5 Dietary estimation

This section describes how dietary intake and food purchase data were processed to generate estimates of daily nutrients consumed and purchased for comparison. First, I describe how dietary estimates are calculated from FFQ data. Next, I describe how comparable estimates are derived from loyalty card purchase records.

Outcomes are expressed in both absolute and energy-adjusted terms and purchase estimates are presented at both the household and individual-level. The outcomes generated are:

- Absolute daily **intake** of energy (Kcal), macronutrients (protein (g), fat (g), saturated fat (g), and sugars (g)), and sodium (mg) at the **individual** level.
- Absolute daily **purchase** of energy (Kcal), macronutrients (protein (g), fat (g), saturated fat (g), and sugars (g)), and sodium (mg) at the **household** level.
- Absolute daily **purchase** of energy (Kcal), macronutrients (protein (g), fat (g), saturated fat (g), and sugars (g)), and sodium (mg) at the **individual** level.
- Energy-adjusted daily **intake** of protein, fat, saturated fat, sugars (% energy) and sodium (mg/Kcal) at the **individual** level.
- Energy-adjusted daily **purchase** of protein, fat, saturated fat, sugars (% energy) and sodium (mg/Kcal) at the **household** level.

2.2.5.1 Daily intake estimates

Daily nutrient intake estimates were provided by the SCG team at the participant level and at the food group level for each participant. Outputs were provided for 51 nutrients (plus water), including energy, macronutrients and micronutrients. For comparison with transaction data, only those nutrients found consistently on a product's back of pack nutritional panel were selected for inclusion in the study (energy (kcal), protein (g), sugars (g), total fats (g), saturated fat (g), sodium (mg)).

Energy-adjusted daily intakes were calculated by taking the absolute daily intake estimate (*NInt*) (in grams) for each macronutrient then multiplying it by the number of calories provided per gram of that nutrient (*Nkcal*). Energy (kcal) values per gram of macronutrient are as follows; protein = 4; fat = 9; saturated fat = 9; sugars = 3.9. The number of calories from a given nutrient is then divided by the absolute estimate for daily calorie intake (*EInt*) and then multiplied by 100 to express each nutrient's calorie contribution as a percentage of daily energy. This is expressed in Equation 1. As sodium does not provide energy, energy-adjusted sodium intake (mg/kcal) is calculated simply by taking the daily sodium intake estimate (mg) and dividing by the daily energy intake (kcal) for each participant.

Equation 1. Calculation for proportion of energy intake from each macronutrient

$$\left(\frac{(NInt * Nkcal)}{EInt}\right) * 100$$

2.2.5.2 Daily purchase estimates

To generate equivalent estimates from purchase data required additional preparatory steps. This section describes these steps to calculate daily nutrient purchase estimates for comparison with daily nutrient intake estimates.

First, the total amount of each nutrient purchased (*PNu*) by each participant at the product level was calculated by multiplying the product's nutrient composition (*Nu*) (per 100g or 100ml) by the amount of product purchased. This was repeated for each of the analysed nutrients and expressed in terms of its given units (energy expressed in kilocalories, macronutrients expressed in grams, and sodium expressed in milligrams). The calculation varied dependent upon whether the product was pre-packaged or sold loose. Packaged goods are sold by volume as stated on the product's packaging, thus the total amount of product purchased by weight is the product's unit weight (*Wt*) multiplied by the number of units purchased (*U*) (Equation 2).

Equation 2. Calculation for total amount of nutrient per product purchased, for pre-packaged products sold by unit

$$PNu = (Nu) * (Wt * U)$$

Unpackaged foods (e.g. unpackaged fruits and vegetables) are sold loose by weight, therefore the total amount of product purchased by weight is the item weight according to the scales at the checkout (*IWt*) (Equation 3).

Equation 3. Calculation for total amount of nutrient per product purchased, for loose products sold by weight

$$PNu = (Nu) * IWt$$

Where *U* had a negative value, this represented items which had been returned by the customer. Items with negative unit values were removed so as not to count towards overall purchased nutrients.

The nutrient composition of some products is expressed per 100g of product, while for others is expressed per 100ml (e.g. beverages, cooking oil, ice cream). In the absence of specific gravity data for the density of products, the assumption that 1ml in volume is equal to 1g in weight was applied, as per the case for pure water. This will mean that the nutrient quantities for some products will have been under-estimated. Product weight/volume information was provided alongside the unit (grams, kilograms, millilitres, litres, or EA (Each/unit)). For consistency, all product weights were converted to grams. Weights expressed as litres or kilograms were multiplied by 1000. For products with missing weight data, or where weight was expressed as the number of units (EA), product weight was imputed as follows. Products with missing weight data were typically those which had missing or unmatched product nutrient composition data (i.e. mostly fresh produce and alcohol).

For a large number of products, the product weight (or volume) was stated in the product description field e.g. “Lager, 500ml”. A formula was applied in Excel to extract all numbers from the product description and consider this the weight in grams. While this method was sufficient for the majority of products it did not work for multipacks (e.g. a multipack of crisps 12 x 25g would be coded as 1225g instead of 300g), products expressed in units other than grams or ml (e.g. Red wine 70cl would be coded as 70g rather than 700g), or products with a number in their product name (e.g. 7-Up 500ml would be coded as 7500g rather than 500g). Extracted weight values were therefore manually checked and corrected against the product description. For products where no weight information was stated in the product description, the weight for a similar item was taken from existing data (e.g. all bottles of wine were coded as 700g and all crisps were assumed to be 25g per bag), or using internet searches.

The absolute amount of each nutrient purchased was summed across all products to give a household-level total for a given timeframe (e.g. baseline, cohort period, study period). This was then divided by the number of days in the timeframe to give the absolute daily household purchase estimates for each nutrient.

Accounting for household composition, absolute daily household purchase estimates were extrapolated to the individual level based on UK daily calorie intake recommendations by age and sex (Table 2.3) (PHE, 2016). Using information provided in the baseline questionnaire, the participant's age and gender were used to ascertain their recommended daily energy intake. If the participant's gender is unknown the mean of the recommendations for males and females is taken. If the participant's age is unknown, they are assumed to be an adult aged between 18 and 64 years. The recommended daily calorie intake for the participant is then divided by the total of calorie intake recommendations for the whole household, based on the number of people reported to be in each age band. Finally, this is multiplied by 100 to express as the percentage of total household nutrients purchased which are expected to be allocated to the study participant. For example, if the participant lives alone they would be allocated 100% of the household nutrients, while if the participant is a 30-year-old woman living with a 30-year-partner and a 3 year old child, she would be allocated 36% of household nutrients ((1928/(1928)+(1*2230)+(1*1197.5))*100). As data was collected on the age of other household members but not their genders, the mean of the recommendations for males and females is taken for all other household members.

Table 2.3 UK recommended daily calorie intakes by age and gender

Data taken from PHE, 2016

	Recommended daily energy intake (kcal)	
Age (years)	Female	Male
0 – 1	698	745
1 – 3	1165	1230
4 – 10	1656	1861
11 – 17	1959	2449
18 – 64	1928	2532
65+	1855	2215

Energy-adjusted purchase estimates were calculated as described in section 2.2.5.1 for energy-adjusted intake estimates. Just one energy-adjusted purchase estimate is required per customer. This is because all nutrients are distributed among household participants according to recommended energy ratios. Thus, the energy-adjusted values for each nutrient at the individual-level would be the same as at the household-level.

2.2.6 Data linkage

As described previously, the STRIDE study utilised data from multiple sources with participant consent; baseline questionnaire, FFQ, transaction records, retailer customer data, retailer product composition data, national food tables, and UK energy intake recommendations. The data relations diagram in Figure 2.11 below (which applies to each cohort) summarises how these data sources were linked and the processing required to generate required outputs for analysis. Each box represents a file (or files) with the colour representing its source.

Orange boxes represent data files provided by the retailer; green boxes represent primary survey data collected for the STRIDE study; grey represents public open access data; blue boxes represent processing steps performed by an R-script; yellow represents output csv files. The variables in each file are listed in the box and unique keys for data linkage are represented by bold type. The data relations diagram depicts the process of data linkage and reduction. Transactions were aggregated at the customer level and transformed from a long format (where each row represents a transaction for a given product made by a particular customer) to a wide format (where each row represents a customer). The file named “Trans_hh_ind.csv” shown by the yellow box at the bottom of the diagram is the final output file which contains the mean daily purchase and intake of each of the nutrients of interest, in absolute and energy-adjusted terms at the household and individual-level for each customer. This is then combined with customer demographic data and used in the analysis. Data processing and analysis was carried out in R studio version 1.4.1106.

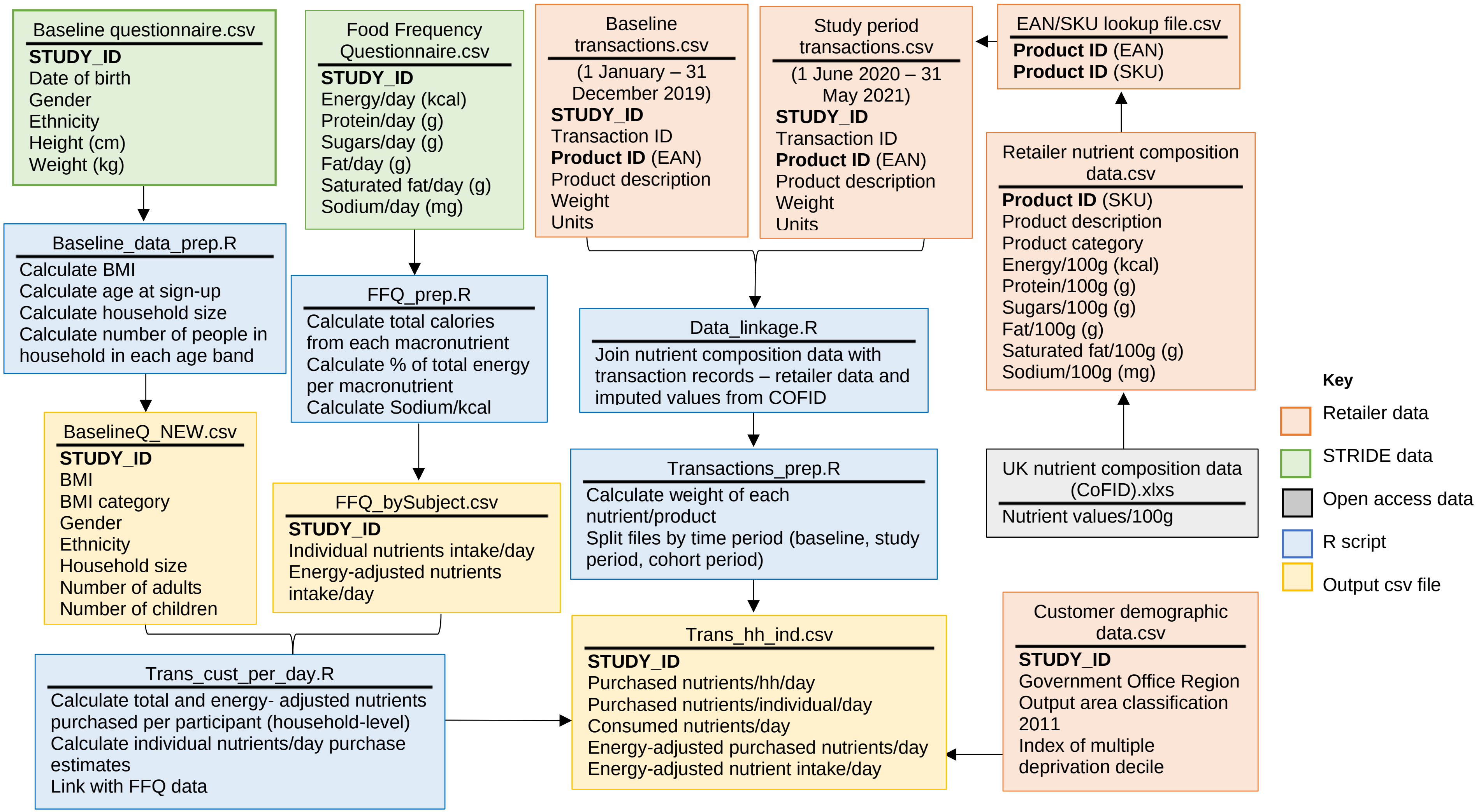

Baseline questionnaire.csv
STUDY_ID
Date of birth
Gender
Ethnicity
Height (cm)
Weight (kg)
Food Frequency Questionnaire.csv
STUDY_ID
Energy/day (kcal)
Protein/day (g)
Sugars/day (g)
Fat/day (g)
Saturated fat/day (g)
Sodium/day (mg)
Baseline transactions.csv
(1 January – 31 December 2019)
STUDY_ID
Transaction ID
Product ID (EAN)
Product description
Weight
Units
Study period transactions.csv
(1 June 2020 – 31 May 2021)
STUDY_ID
Transaction ID
Product ID (EAN)
Product description
Weight
Units
EAN/SKU lookup file.csv
Product ID (EAN)
Product ID (SKU)
Retailer nutrient composition data.csv
Product ID (SKU)
Product description
Product category
Energy/100g (kcal)
Protein/100g (g)
Sugars/100g (g)
Fat/100g (g)
Saturated fat/100g (g)
Sodium/100g (mg)
Baseline_data_prep.R
Calculate BMI
Calculate age at sign-up
Calculate household size
Calculate number of people in household in each age band
FFQ_prep.R
Calculate total calories from each macronutrient
Calculate % of total energy per macronutrient
Calculate Sodium/kcal
Data_linkage.R
Join nutrient composition data with transaction records – retailer data and imputed values from COFID
BaselineQ_NEW.csv
STUDY_ID
BMI
BMI category
Gender
Ethnicity
Household size
Number of adults
Number of children
FFQ_bySubject.csv
STUDY_ID
Individual nutrients intake/day
Energy-adjusted nutrients intake/day
Transactions_prep.R
Calculate weight of each nutrient/product
Split files by time period (baseline, study period, cohort period)
UK nutrient composition data (CoFID).xlxs
Nutrient values/100g
Key
Retailer data
STRIDE data
Open access data
R script
Output csv file
Trans_cust_per_day.R
Calculate total and energy- adjusted nutrients purchased per participant (household-level)
Calculate individual nutrients/day purchase estimates
Link with FFQ data
Trans_hh_ind.csv
STUDY_ID
Purchased nutrients/hh/day
Purchased nutrients/individual/day
Consumed nutrients/day
Energy-adjusted purchased nutrients/day
Energy-adjusted nutrient intake/day
Customer demographic data.csv
STUDY_ID
Government Office Region
Output area classification 2011
Index of multiple deprivation decile

Figure 2.11 STRIDE data relations diagram

2.2.7 Impacts of COVID-19

The STRIDE study took place during 2020 and 2021 in the context of the COVID-19 global pandemic. The pilot phase was originally designed to launch in January 2020, with completion of the FFQ in March 2020 giving a data coverage between January and March 2020 for pilot participants. However, commencement was delayed due to retailer resource constraints as they worked hard to manage the increased pressures placed upon supermarkets due to panic buying at the beginning of 2020. The pilot phase launched in March, three months later than planned, giving a data coverage period from March – May 2020. As a result, there is a gap in data coverage by transaction data from the end of the baseline transaction period (1 January 2019 – 31 December 2019) to the beginning of the pilot phase (1 March 2020).

Incidentally, the launch of the pilot phase coincided with the first UK National Lockdown which came into force on 26 March 2020 and remained in place until 23 June (Institute for Government., 2021). Cohorts 1 – 3 all spanned periods which were subject to varying local and national lockdown restriction measures across the UK. To capture the impact of the COVID-19 pandemic on food purchasing and dietary intake, the baseline questionnaire was updated after the pilot phase to include questions about the effect of lockdown restrictions. This required an update to the original ethics form. The full list of baseline questions is given in Appendix B.5. It is not clear how the pandemic might have affected recruitment and data collection during the STRIDE study. This, and the potential impact on dietary coverage are discussed in further detail in the final discussion chapter (Chapter 6).

2.3 Data and methods chapter summary

In this chapter, I have given a detailed overview of the data preparation and methods which contribute to the papers in this thesis. The additional detail in this chapter, with a particular focus on the use of secondary retail data and the design of the STRIDE study, aims to provide the reader with a more in-depth understanding of the significant partnership investment, data hosting resource and up-front data preparation which is required to repurpose supermarket transaction records for population dietary monitoring. The proceeding three chapters are made up of papers representing the substantive research contribution of the thesis. First, Chapter 3 presents a systematic review of the literature, critiquing how supermarket transactions have contributed to dietary research to date and what remains to be investigated. Next, Chapter 4 provides a case study for the use of supermarket transaction records to uncover small area dietary patterns. This will be followed by findings from the STRIDE study in Chapter 5, which addresses the validity of supermarket transactions against self-reported intake.

References

Arribas-Bel, D. 2014. Accidental, Open and Everywhere: Emerging Data Sources for the Understanding of Cities. *Applied Geography.* **49**, pp.45–53.

Bandy, L., Adhikari, V., Jebb, S. and Rayner, M. 2019. The use of commercial food purchase data for public health nutrition research: A systematic review. *PLOS ONE.* **14**(1), pe0210192.

Baty, S., Butchers, M., Moss, M. 2020. FeedUK – building resilience by digitising the food system. *Food Science and Technology.* **34**(4), pp.46-49.

Bland, M.J., Altman, D.G. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* **i**, pp.307-310.

Bonevski, B., Randell, M., Paul, C., Chapman, K., Twyman, L., Bryant, J., Brozek, I. and Hughes, C. 2014. Reaching the hard-to-reach: a systematic review of strategies for improving health and medical research with socially disadvantaged groups. *BMC Medical Research Methodology.* **14**(1), p42.

Carter, M., Albar, S., Morris, M., Mulla, U., Hancock, N., Evans, C., Alwan, N., Greenwood, D., Hardie, L., Frost, G., Wark, P. and Cade, J. 2015. Development of a UK Online 24-h Dietary Assessment Tool: myfood24. *Nutrients.* **7**(6), p4016.

Casas, R., Castro-Barquero, S., Estruch, R. and Sacanella, E. 2018. Nutrition and Cardiovascular Health. *International journal of molecular sciences.* **19**(12), p3988.

Chen, G., Jia, W., Zhao, Y., Mao, Z.-H., Lo, B., Anderson, A.K., Frost, G., Jobarteh, M.L., McCrory, M.A., Sazonov, E., Steiner-Asiedu, M., Ansong, R.S., Baranowski, T., Burke, L. and Sun, M. 2021. Food/Non-Food Classification of Real-Life Egocentric Images in Low- and Middle-Income Countries Based on Image Tagging Features. *Frontiers in Artificial Intelligence.* **4**.

Clark, S.D., Shute, B., Jenneson, V., Rains, T., Birkin, M. and Morris, M.A. 2021. Dietary Patterns Derived from UK Supermarket Transaction Data with Nutrient and Socioeconomic Profiles. *Nutrients.* **13**(5), p1481.

de la Hunty, A., Buttriss, J., Draper, J., Roche, H., Levey, G., Florescu, A., Penfold, N. and Frost, G. 2021. UK Nutrition Research Partnership (NRP) workshop: Forum on advancing dietary intake assessment. *Nutrition Bulletin.* **46**(2), pp.228-237.

Department of Health. 2017. *Technical guidance on nutrition labelling.* London, UK: UK Government. [Online] [Accessed 13.06.2018]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/595961/Nutrition_Technical_Guidance.pdf

Desai, T., Ritchie, F., Welpton, R. 2016. *The Five Safes: designing data access for research*. Working papers in Economics. University of the West of England, Bristol. 1601.

Dietary Assessment Primer. 2022. *Glossary of Key Terms.* [Online]. [Accessed 12.01.2022]. Available from:

https://dietassessmentprimer.cancer.gov/glossary.html#concentration_biomarker

DOH. 2015. *Public Health Responsibility Deal.* [Online]. [Accessed 13.06.2018]. Available from: http://webarchive.nationalarchives.gov.uk/20180201175643/https://responsibilitydeal.dh.gov.uk/

Einav, L., Leibtag, Ephraim., Nevo, Aviv. 2008. *On the Accuracy of Nielsen Homescan Data.* U.S. Department of Agriculture.

Eyles, H., Jiang, Y. and Ni Mhurchu, C. 2010. Use of Household Supermarket Sales Data to Estimate Nutrient Intakes: A Comparison with Repeat 24-Hour Dietary Recalls. *Journal of the American Dietetic Association.* **110**(1), pp.106-110.

Food Foundation. 2020a. *Peas Please Progress Report 2020.* UK. [Online]. [Accessed 14.09.2021]. Available from: https://foodfoundation.org.uk/publication/peas-please-progress-report-2020/

Food Foundation. 2020b. *Peas Please. Reviewing the evidence: what can retailers do to increase sales of fruit and veg.* London, UK. [Online]. [Accessed 14.09.2021]. Available from: https://foodfoundation.org.uk/retailer-toolkit/research/

Gale, C., Singleton, A., Bates, A. and Longley, P. 2016. Creating the 2011 area classification for output areas (2011 OAC). *Journal of Spatial Information Science.* **12**.

Gemming, L., Utter, J. and Ni Mhurchu, C. 2015. Image-assisted dietary assessment: a systematic review of the evidence. *Journal of the Academy of Nutrition and Dietetics.* **115**(1), pp.64-77.

GOV.UK. 2015a. *75 Years of Family Food.* [Online]. [Accessed 12.01.2022]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/597666/FF75Timeline-09mar17.pdf

GOV.UK. 2015b. *English indices of deprivation 2015.* [Online]. [Accessed 12.01.2022]. Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015

GOV.UK. 2018. *Regional ethnic diversity.* [Online]. [Accessed 12.01.2022]. Available from: https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/national-and-regional-populations/regional-ethnic-diversity/latest

GOV.UK. 2020a. *Collection: Family food statistics. Annual statistics about food and drink purchases in the UK.* [Online]. [Accessed 12.01.2022]. Available from: https://www.gov.uk/government/collections/family-food-statistics

GOV.UK. 2020b. *Consultation outcome: Restricting promotions of products high in fat, sugar and salt by location and by price: government response to public consultation.* [Online]. [Accessed 08/01/2021]. Available from: https://www.gov.uk/government/consultations/restricting-promotions-of-food-and-drink-that-is-high-in-fat-sugar-and-salt/outcome/restricting-promotions-of-products-high-in-fat-sugar-and-salt-by-location-and-by-price-government-response-to-public-consultation

Griffith, R., Jenneson, V., James, J. and Taylor, A. 2021. *The impact of a tax on added sugar and salt. IFS working paper W21/21.*

Griffith, R. and O'Connell, M. 2009. The Use of Scanner Data for Research into Nutrition. *Fiscal Studies.* **30**(3-4), pp.339-365.

Gutierrez, L., Folch, A., Rojas, M., Cantero, J.L., Atienza, M., Folch, J., Camins, A., Ruiz, A., Papandreou, C. and Bulló, M. 2021. Effects of Nutrition on Cognitive Function in Adults with or without Cognitive Impairment: A Systematic Review of Randomized Controlled Clinical Trials. *Nutrients.* **13**(11), p3728.

Hawkes, C., Smith, T.G., Jewell, J., Wardle, J., Hammond, R.A., Friel, S., Thow, A.M. and Kain, J. 2015. Smart food policies for obesity prevention. *The Lancet.* **385**(9985), pp.2410-2421.

Heald, C.L., Bolton-Smith, C., Ritchie, M.R., Morton, M.S. and Alexander, F.E. 2006. Phyto-oestrogen intake in Scottish men: use of serum to validate a self-administered food-frequency questionnaire in older men. *Eur J Clin Nutr.* **60**(1), pp.129-135.

Hebert, J.R., Clemow, L., Pbert, L., Ockene, I.S. and Ockene, J.K. 1995. Social Desirability Bias in Dietary Self-Report May Compromise the Validity of Dietary Intake Measures. *International Journal of Epidemiology.* **24**(2), pp.389-398.

Hebert, J.R., Ma, Y., Clemow, L., Ockene, I.S., Saperia, G., Stanek, E.J., III, Merriam, P.A. and Ockene, J.K. 1997. Gender Differences in Social Desirability and Social Approval Bias in Dietary Self-report. *American Journal of Epidemiology.* **146**(12), pp.1046-1055.

HMRC. 2018. *Check if your drink is liable for the Soft Drink Industry Levy.* [Online]. [Accessed 13.08.19]. Available from: https://www.gov.uk/guidance/check-if-your-drink-is-liable-for-the-soft-drinks-industry-levy

Hollis, J.L., Craig, L.C., Whybrow, S., Clark, H., Kyle, J.A. and McNeill, G. 2017. Assessing the relative validity of the Scottish Collaborative Group FFQ for measuring dietary intake in adults. *Public Health Nutr.* **20**(3), pp.449-455.

IGD. 2021a. *IGD's Healthy and Sustainable Diets Project Group.* [Online]. [Accessed 24.01.2022]. Available from: https://www.igd.com/charitable-impact/healthy-eating/content-library/article/t/igds-healthy-and-sustainable-diets-project-group/i/28089

IGD. 2021b. *Healthy, sustainable diets: driving change.* [Online]. [Accessed 24.01.2022]. Available from: https://www.igd.com/social-impact/health/shifting-consumer-behaviour/article-viewer/t/healthy-sustainable-diets-driving-change/i/29113

Institute for Government. 2021. *Timeline of UK coronavirus lockdowns, March 2020 to March 2021.* [Online]. [Accessed 24.01.2022]. Available from: https://www.instituteforgovernment.org.uk/sites/default/files/timeline-lockdown-web.pdf

Jenneson, V., Clarke, G.P., Greenwood, D.C., Shute, B., Tempest, B., Rains, T. and Morris, M.A. 2022. Exploring the Geographic Variation in Fruit and Vegetable Purchasing Behaviour Using Supermarket Transaction Data. *Nutrients.* **14**(1), p177.

Jenneson, V., Morris, M., Greenwood, D. and Clarke, G. 2020. *STRIDE Study (Supermarket Transaction Records In Dietary Evaluation) protocol*. Open Science Framework. Available from: https://doi.org/10.17605/OSF.IO/VUKTQ

Jenneson, V., Pontin, F., Greenwood, D.C., Clarke, G.P. and Morris, M.A. 2021. A systematic review of supermarket automated electronic sales data for population dietary surveillance. *Nutrition Reviews.*

Jia, X., Craig, L.C.A., Aucott, L.S., Milne, A.C. and McNeill, G. 2008. Repeatability and validity of a food frequency questionnaire in free-living older people in relation to cognitive function. *The Journal of Nutrition, Health & Aging.* **12**(10), pp.735-741.

LIDA. 2021. *LIDA announces partnership with Sainsbury's.* [Online]. [Accessed 24.01.2022]. Available from: https://lida.leeds.ac.uk/news/sainsburys-partnership/

Liu, B., Young, H., Crowe, F.L., Benson, V.S., Spencer, E.A., Key, T.J., Appleby, P.N. and Beral, V. 2011. Development and evaluation of the Oxford WebQ, a low-cost, web-based method for assessment of previous 24 h dietary intakes in large-scale prospective studies. *Public Health Nutrition.* **14**(11), pp.1998-2005.

Masson, L.F., McNeill, G., Tomany, J.O., Simpson, J.A., Peace, H.S., Wei, L., Grubb, D.A. and Bolton-Smith, C. 2003. Statistical approaches for assessing the relative validity of a food-frequency questionnaire: use of correlation coefficients and the kappa statistic. *Public Health Nutrition.* **6**(3), pp.313-321.

McCambridge, J., Witton, J. and Elbourne, D.R. 2014. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *Journal of clinical epidemiology.* **67**(3), pp.267-277. *diet: noun (1). Definition of diet (Entry 1 of 4)*. 2022. s.v.

Mohd-Shukri, N.A., Bolton, J.L., Norman, J.E., Walker, B.R. and Reynolds, R.M. 2013. Evaluation of an FFQ to assess total energy and nutrient intakes in severely obese pregnant women. *Public Health Nutrition.* **16**(8), pp.1427-1435.

Morris, M., Glaser, A. and Iles-Smith, H. 2018a. *Study protocol: a survey of public opinion on the acceptability of linking electronic health records with loyalty card (supermarket) and liefestyle app data for research.* LifeInfo Survey: What do you think about researchers using your lifestyle information? : Open Science Framework.,.

Morris, M., Wilkins, E., Timmins, K., Bryant, M., Birkin, M. and Griffiths, C. 2018b. Can big data solve a big problem? Reporting the obesity data landscape in line with the Foresight obesity system map. *Int J Obes (Lond).* **42**(12), pp.1963-1976.

MRC. 2017. *Review of Nutrition and Human Health Research.* [Online]. Available from: https://mrc.ukri.org/documents/pdf/review-of-nutrition-and-human-health/

National Food Strategy. 2021. *The Plan: National Food Strategy, Independent Review.* London, UK.

Nestle, M. 2003. *Food Politics: How the Food Industry Influences Nutrition and Health.* California: University of California Press.

Neuenschwander, M., Ballon, A., Weber, K.S., Norat, T., Aune, D., Schwingshackl, L. and Schlesinger, S. 2019. Role of diet in type 2 diabetes incidence: umbrella review of meta-analyses of prospective observational studies. *BMJ.* **366**, pl2368.

Nutritools. 2019. *Nutritools, Dietary Assessment Tools, Tool Library.* [Online]. [Accessed 25.01.2022]. Available from: https://www.nutritools.org/tools

OECD. 2019. *The Heavy Burden of Obesity.*
Office for National Statistics. 2017. *Living costs and food survey: user guidance and technical information for the Living Costs and Food Survey.* [Online]. [Accessed 12.01.2022]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/methodologies/livingcostsandfoodsurvey

Office for National Statistics. 2021. Family Spending: Workbook 1 - detailed expenditure and trends. Table A2. London, UK: *Office for National Statistics*. [Online]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/expenditure/datasets/familyspendingworkbook1detailedexpenditureandtrends

ONS. 2017. *Output areas: Introduction to Output Areas - the building block of Census geography.* [Online]. [Accessed 23.12.17]. Available from: https://www.ons.gov.uk/census/2001censusandearlier/dataandproducts/outputgeography/outputareas

Oxford English Dictionary. 2021. *"diet, n.1".* Oxford University Press.

Pell, D., Mytton, O., Penney, T.L., Briggs, A., Cummins, S., Penn-Jones, C., Rayner, M., Rutter, H., Scarborough, P., Sharp, S.J., Smith, R.D., White, M. and Adams, J. 2021. Changes in soft drinks purchased by British households associated with the UK soft drinks industry levy: controlled interrupted time series analysis. *BMJ.* **372**, pn254.

PHE. 2016. *Government Dietary Recommendations. Government recommendations for energy and nutrients for males and females aged 1 - 18 years and 19+ years.* London, UK: UK Government.

PHE. 2020. McCance and Widdowson's composition of foods integrated dataset. London, UK: *gov.uk*. [Online]. [Accessed 21.01.2022]. Available from: https://www.gov.uk/government/publications/composition-of-foods-integrated-dataset-cofid

PHE. 2021. *Evaluation of changes in the dietary methodology in the National Diet and Nutrition Survey Rolling Programme from Year 12 (2019 to 2020).* London, UK: Public Health England.,.

Public Health England. 2015. *Composition of Foods Integrated Dataset (CoFID). McCance and Widdowson's composition of foods: old foods.* [Online]. [Accessed 25.01.2022]. Available from: https://www.gov.uk/government/publications/composition-of-foods-integrated-dataset-cofid

Qiu, J., Lo, F.P.W., Jiang, S., Tsai, Y.Y., Sun, Y. and Lo, B. 2021. Counting Bites and Recognizing Consumed Food from Videos for Passive Dietary Monitoring. *IEEE Journal of Biomedical and Health Informatics.* **25**(5), pp.1471-1482.

Ransley, J.K., Donnelly, J.K., Khara, T.N., Botham, H., Arnot, H., Greenwood, D.C. and Cade, J.E. 2001. The use of supermarket till receipts to determine the fat and energy intake in a UK population. *Public Health Nutrition.* **4**(6), pp.1279-1286.

Rehm, J., Kilian, C., Rovira, P., Shield, K.D. and Manthey, J. 2021. The elusiveness of representativeness in general population surveys for alcohol. *Drug and Alcohol Review.* **40**(2), pp.161-165.

Russell, W.R., Baka, A., Björck, I., Delzenne, N., Gao, D., Griffiths, H.R., Hadjilucas, E., Juvonen, K., Lahtinen, S., Lansink, M., Loon, L.V., Mykkänen, H., östman, E., Riccardi, G., Vinoy, S. and Weickert, M.O. 2016. Impact of Diet Composition on Blood Glucose Regulation. *Critical Reviews in Food Science and Nutrition.* **56**(4), pp.541-590.

Simpson, E., Bradley, J., Poliakov, I., Jackson, D., Olivier, P., Adamson, A.J. and Foster, E. 2017. Iterative Development of an Online Dietary Recall Tool: INTAKE24. *Nutrients.* **9**(2), p118.

Springmann, M., Wiebe, K., Mason-D'Croz, D., Sulser, T.B., Rayner, M. and Scarborough, P. 2018. Health and nutritional aspects of sustainable diet strategies and their association with environmental impacts: a global modelling analysis with country-level detail. *The Lancet Planetary Health.* **2**(10), pp.e451-e461.

Statista. 2021. *Market share of grocery stores in Great Britain from January 2017 to May 2021.* [Online]. [Accessed 24.01.2022]. Available from: https://www.statista.com/statistics/280208/grocery-market-share-in-the-united-kingdom-uk/

The Consumer Goods Forum. 2020. *Better Lives Through Better Business.* UK. [Online]. [Accessed 24.01.2022]. Available from: https://www.theconsumergoodsforum.com/health-wellness/healthier-lives/about/mission/

The Consumer Goods Forum. 2021. *Mission: empower people to lead healthier lives while creating shared value for business and communities.* [Online]. [Accessed 24.01.2022]. Available from: https://www.theconsumergoodsforum.com/health-wellness/healthier-lives/about/mission/

Theis, D.R.Z. and White, M. 2021. Is Obesity Policy in England Fit for Purpose? Analysis of Government Strategies and Policies, 1992–2020. *The Milbank Quarterly.* **99**(1), pp.126-170.

Valdes, A.M., Walter, J., Segal, E. and Spector, T.D. 2018. Role of the gut microbiota in nutrition and health. *BMJ.* **361**, pk2179.

Vepsäläinen, H., Nevalainen, J., Kinnunen, S., Itkonen, S.T., Meinilä, J., Männistö, S., Uusitalo, L., Fogelholm, M. and Erkkola, M. 2021. Do we eat what we buy? Relative validity of grocery purchase data as an indicator of food consumption in the LoCard study. *British Journal of Nutrition.* pp.1-24.

Wansink, B. 2010. From mindless eating to mindlessly eating better. *Physiology & Behavior.* **100**(5), pp.454-463.

Wark, P.A., Hardie, L.J., Frost, G.S., Alwan, N.A., Carter, M., Elliott, P., Ford, H.E., Hancock, N., Morris, M.A., Mulla, U.Z., Noorwali, E.A., Petropoulou, K., Murphy, D., Potter, G.D.M., Riboli, E., Greenwood, D.C. and Cade, J.E. 2018. Validity of an online 24-h recall tool (myfood24) for dietary assessment in population studies: comparison with biomarkers and standard interviews. *BMC Med.* **16**(1), p136.

WCRF. 2018. *Diet, Nutrition, Physical Activity and Cancer: a Global Perspective.* World Cancer Research Fund/American Institute for Cancer Research.

Wehling, H. and Lusher, J. 2019. People with a body mass index ⩾30 under-report their dietary intake: A systematic review. *Journal of Health Psychology.* **24**(14), pp.2042-2059.

WHO. 2021. *Report of the technical consultation on measuring healthy diets: concepts, methods and metrics. Virtual meeting, 18 - 20 May 2021.* Geneva, Switzerland: World Health Organisation.

Willett, W., Rockström, J., Loken, B., Springmann, M., Lang, T., Vermeulen, S., Garnett, T., Tilman, D., DeClerck, F., Wood, A., Jonell, M., Clark, M., Gordon, L.J., Fanzo, J., Hawkes, C., Zurayk, R., Rivera, J.A., De Vries, W., Majele Sibanda, L., Afshin, A., Chaudhary, A., Herrero, M., Agustina, R., Branca, F., Lartey, A., Fan, S., Crona, B., Fox, E., Bignet, V., Troell, M., Lindahl, T., Singh, S., Cornell, S.E., Srinath Reddy, K., Narain, S., Nishtar, S. and Murray, C.J.L. 2019. Food in the Anthropocene: the EAT&#x2013;<em>Lancet</em> Commission on healthy diets from sustainable food systems. *The Lancet.* **393**(10170), pp.447-492.

World Health Organization, F., Agriculture Organization of the United Nations. and Joint WHO FAO Expert Consultation on Diet Nutrition and the Prevention of Chronic Diseases. 2003. *Diet, nutrition, and the prevention of chronic diseases : report of a joint WHO/FAO expert consultation.* Geneva: World Health Organization.

Part 2

Included publications

Chapter 3
A systematic review of automated electronic supermarket sales data for population dietary surveillance

Victoria L Jenneson, Francesca Pontin, Darren C Greenwood, Graham P Clarke, Michelle A Morris

Nutrition Reviews. **2021**. nuab089, https://doi.org/10.1093/nutrit/nuab089

Keywords: dietary assessment, dietary surveillance, methods, supermarket transactions
Topic: diet, food, dietary assessment, nutrients, surveillance, medical, supermarkets
Issue Section: Special article

Abstract

Context

Most dietary assessment methods are limited by self-report biases, how long they take for participants to complete, and cost of time for dietitians to extract content. Electronically recorded, supermarket-obtained transactions are an objective measure of food purchases, with reduced bias and improved timeliness and scale.

Objective

The use, breadth, context, and utility of electronic purchase records for dietary research is assessed and discussed in this systematic review.

Data sources

Four electronic databases (MEDLINE, EMBASE, PsycINFO, Global Health) were searched. Included studies used electronically recorded supermarket transactions to investigate the diet of healthy, free-living adults.

Data extraction

Searches identified 3422 articles, of which 145 full texts were retrieved and 72 met inclusion criteria. Study quality was assessed using the National Institutes of Health Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies.

Data analysis

Purchase records were used in observational studies, policy evaluations, and experimental designs. Nutrition outcomes included dietary patterns, nutrients, and food category sales. Transactions were linked to nutrient data from retailers, commercial data sources, and national food composition databases.

Conclusion

Electronic sales data have the potential to transform dietary assessment and worldwide understanding of dietary behaviour. Validation studies are warranted to understand limits to agreement and extrapolation to individual-level diets.

Systematic Review Registration

PROSPERO registration no. CRD42018103470

3.1 Introduction

Population dietary surveillance is important for understanding temporal changes and variation between subgroups. This contributes to the epidemiological understanding of diet-related diseases (Mooney and Pejaver, 2018) and enables targeting and evaluation of public health policy interventions. Current approaches to population dietary surveillance, including national surveys, rely heavily on self-reported measures of intake and food purchases. Due to their expense, surveys are restrictive in size and geographic coverage. Self-reported dietary measures are often criticised for their introduction of recall and reporting biases on the part of study participants, and possible coding errors by researchers (Buttriss et al., 2017), resulting in a tendency to underestimate intake (Serra-Majem et al., 2003). Moreover, it is not possible for national surveys to collect data continuously or in real-time which limits their utility.

Supermarkets dominate household food supply in high income countries. Thus, supermarket purchase records may offer insight into diets in high (and middle) income settings. Early work using paper cash register receipts highlighted the feasibility of supermarket purchase data to contribute to population dietary surveillance (Ransley et al., 2001). Whilst promising, the paper-based nature of data collection limited scale and timeliness and was reliant upon manual researcher coding (Ransley et al., 2003; Ransley et al., 2001). Recent advancements in computational storage and power preceded a movement for repurposing commercial ‘big data’ sources (Laney, 2013; Kitchin and McArdle, 2016) to address public health and social science questions (Mooney and Pejaver, 2018; Arribas-Bel, 2014; Timmins et al., 2018). Electronic supermarket transaction records, generated as a by-product of daily activity, build upon the earlier foundations of paper-based receipt collection (Ransley et al., 2001). However, they capture purchases rather than consumption and exclude foods eaten out of the home or purchased or obtained elsewhere. Exploration of the utility of supermarket transaction records in nutrition research is therefore warranted.

A previous review of both paper-based and electronically captured transaction records, suggested that supermarket data could contribute to dietary research in seven key areas; 1) dietary patterns, 2) longitudinal analysis, 3) nutrient availability, 4) validation of self-report, 5) identifying predictors of healthy food choices, 6) evaluating intervention effectiveness and 7) exploring associations between diet and health outcomes (Tin et al., 2007). Electronically captured purchase data could offer benefits over paper-based methods, as a more cost-effective, low burden tool for monitoring household dietary purchases, longitudinally and at scale (Tin et al., 2007). However, the review emphasised that challenges related to data linkage and data sharing must be overcome. Furthermore, there is a need for robust analytical methods and to establish correction factors to account for differences between food purchases and consumption (Tin et al., 2007).

Similarly, a recent systematic review by Bandy et al. (2018) highlighted the utility of purchase data from households participating in commercial market research panels as a source of dietary surveillance information. Market research panel data benefits from a large population, coverage of retailers, and temporal granularity. This makes it a useful tool for evaluating national policies (Bandy et al., 2018). However, as with survey methods, data collection is burdensome for participants, and not without reporting biases (Einav, 2008). Furthermore, the cost of access is potentially prohibitive to use by many researchers. In acknowledgement of these limitations, Bandy et al. (2018) called for a review of electronically captured sales data gained directly from supermarket retailers, as an alternative objective source of food purchase data. This review aims to address this gap, and to provide an update to the previous review by Tin et al. (2007).

The review will synthesise existing studies to understand the utility of electronically captured supermarket purchase records in dietary research and offer a clearer understanding of benefits and methodological challenges faced. This will facilitate methodological innovation in dietary assessment, in turn contributing to a better understanding of dietary behaviours worldwide. Thus, this systematic review uses a narrative approach to address the following questions:

1) What types of studies use electronic supermarket transaction records to assess diet-related behaviours in adults?
2) Is supermarket transaction data a valid dietary assessment measure?
3) What sources of nutrient data did the studies use?
4) What nutritional outcomes did the studies report?

3.2 Methods

This review is reported in line with guidance on the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) (Supplementary Table S1). The protocol for this review was published in advance on PROSPERO (CRD42018103470) (Jenneson et al. 2018)

3.2.1 Search strategy

Four electronic journal databases were searched; MEDLINE, EMBASE, PsycINFO and Global Health, for papers published in the English language using MeSH subject headings and keywords relating to diet or nutritional assessment and purchase data (e.g. diet, nutrition assessment, grocery store, purchase, loyalty card). An example search strategy can be found in Supplementary Table S2. Citations were imported into EndNote reference manager and titles and abstracts independently screened against inclusion criteria by two reviewers (VJ and FP). Full texts were requested for all eligible titles and independently screened by VJ and FP. A third reviewer (MM) was available throughout the screening process to resolve any disagreements. Reference lists of identified papers and hand searching were used to identify additional papers for inclusion.

3.2.2 Study selection and data extraction

Studies of any design using electronically captured supermarket sales data to assess dietary outcomes (any measure) were included in this review. Studies using paper cash register receipts or purchase data from market research panels are excluded. Studies measuring non-nutritional aspects of diet (such as organic, fair trade etc) are also excluded. Purchase data may be captured at the individual or household level in the general healthy free-living population and purchases should be carried out by adults aged 18 and above. The full eligibility criteria are described in Table 3.1.

Table 3.1 PICOS criteria for inclusion and exclusion of studies.

	Inclusion criteria
Participants	- Adults ≥18 years old (not purchases made exclusively by children, although children may be part of the household) - Individuals or households - Healthy (disease status unknown) - Free-living
Interventions	- Electronically captured supermarket purchase records - Purchases made at the individual or household level - Not purchases made by organisations or at a national level (e.g. food balance sheets)
Comparisons	- N/A
Outcomes	- Volume or value-based food and/or beverage purchases - Purchased macro/micro-nutrient quantity

	- Nutritional quality of purchased products (e.g. nutrient profile) - Dietary pattern derived from purchased products - Electronically captured purchase records derived from supermarkets - Not paper-based cash register receipts - Not self-reported purchases - Not purchase records collected by market research panels - Not purchases made in laboratory-based experimental studies - Not non-nutritional outcomes e.g. fair trade, organic, food safety
Study Design	- Randomised Controlled Trial - Cohort - Cross-sectional - Quasi-experimental - Not reviews

The two reviewers (VJ and FP) piloted a data extraction form, adapted from The Cochrane Public Health Group Data Extraction and Assessment Template (Cochrane Public Health Group., 2011), for two papers. This was accepted and can be found in Supplementary Table S3. The data extraction form incorporates two of the key elements identified in the BEE COAST framework for reporting big data for an obesity research context (Morris, et al., 2018); description of the original data purpose and aggregation level. Data extraction was carried out by the lead reviewer (VJ).

Dietary outcomes for inclusion are; quantity of sales at a product or category level (expressed as expenditure or volume), purchased macro- and micro-nutrients, and dietary patterns.

3.2.3 Quality assessment

The NIH Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies (NIH, 2017) was used for risk of bias assessment. Studies were assessed by the lead reviewer (VJ) against questions on 13 domain areas, which were answered ‘Yes’, ‘No’, ‘Not Applicable’ or ‘Not reported/Could not determine’. The tool does not use a points system to generate an overall quality score. Instead, the answers to each of these domains contributed to an overall judgement made on the quality of study design and reporting; “Good”, “Fair” or “Poor”. Studies rated ‘Good’ had a maximum of three domains which were not answered ‘Yes’. ‘Validity of outcomes’ and ‘adjustment for confounders’ were considered the most important domains determining classification of ‘Poor’ study quality.

3.2.4 Data synthesis

Due to the variability in study outcomes and methodologies, it was not possible to quantitatively synthesise the study data. Instead, a systematic narrative synthesis approach was used to explore findings and methods thematically, in line with the proposed research questions. Guidance on Narrative Synthesis in Systematic Reviews by the ESRC Methods Programme was followed (Popay, 2006).

3.3 Results

3.3.1 Search results

As the PRISMA flow chart in Figure 3.1 shows, searches returned a total of 3,422 papers published between 1996 and June 2020. From these, 1,862 duplicate records were removed, a further 1,415 papers were removed after dual screening of titles and abstracts (Figure 3.1). Of the remaining 145 papers that underwent full text screening, 62 met the eligibility criteria. These were supplemented by 10 additional papers, which were identified from the reference lists of included studies, giving a total of 72 papers. A detailed summary of papers included in this review can be found Supplementary Table S4.

3.3.2 Study characteristics

Routinely collected electronic sales data were used to monitor dietary outcomes in 72 papers (53 unique studies) across 14 high income countries between 1996 and 2019 (Table 3.2). Following initial interest in the late 1990s, publication of studies using electronic supermarket purchases declined, but has been rising more recently. This reflects both the increasing availability of data and interest in its utility for dietary research.

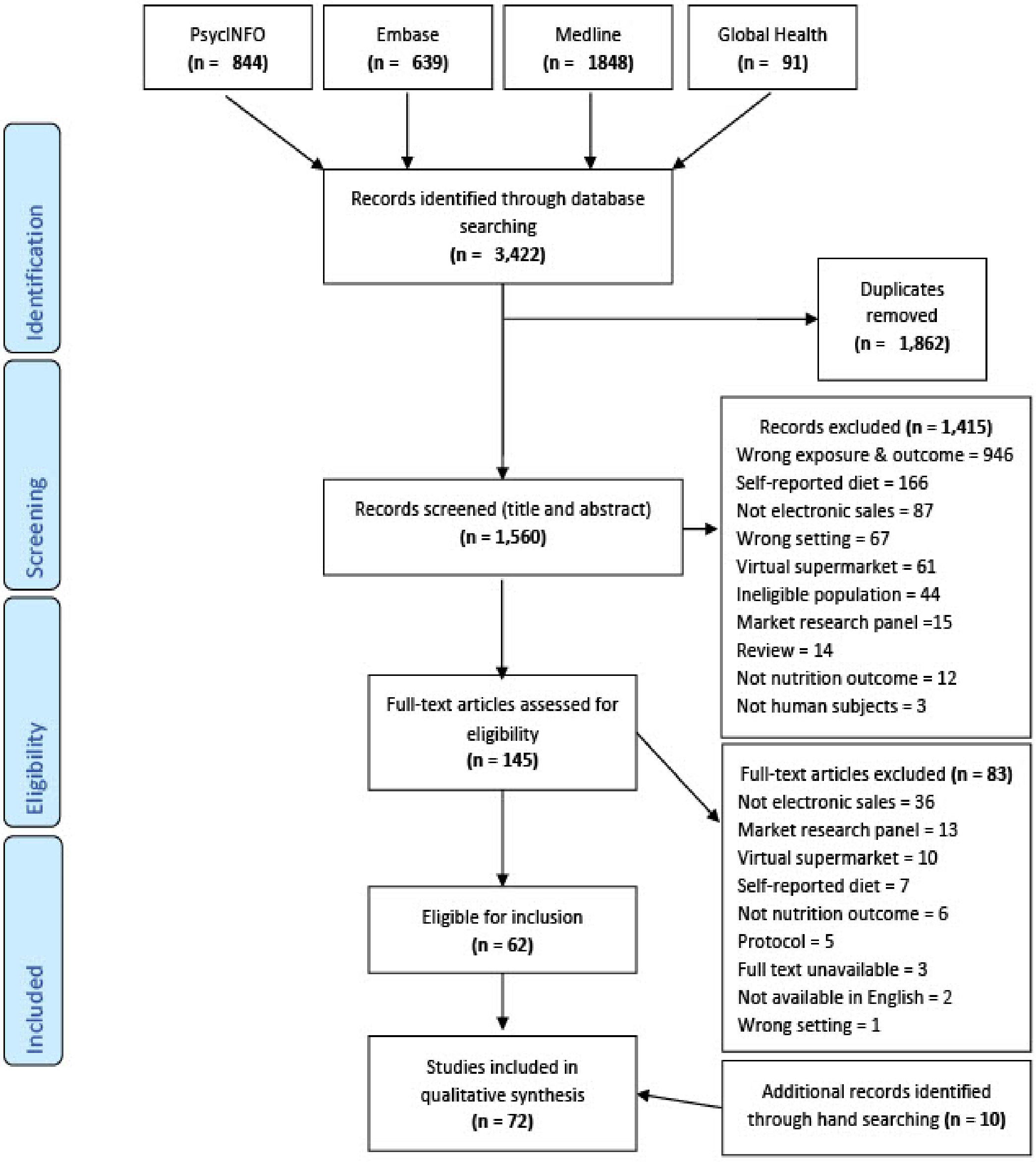

PsycINFO (n = 844)
Embase (n = 639)
Medline (n = 1848)
Global Health (n = 91)
Identification
Records identified through database searching (n = 3,422)
Duplicates removed (n = 1,862)
Screening
Records screened (title and abstract) (n = 1,560)
Records excluded (n = 1,415)
Wrong exposure & outcome = 946
Self-reported diet = 166
Not electronic sales = 87
Wrong setting = 67
Virtual supermarket = 61
Ineligible population = 44
Market research panel =15
Review = 14
Not nutrition outcome = 12
Not human subjects = 3
Eligibility
Full-text articles assessed for eligibility (n = 145)
Full-text articles excluded (n = 83)
Not electronic sales = 36
Market research panel = 13
Virtual supermarket = 10
Self-reported diet = 7
Not nutrition outcome = 6
Protocol = 5
Full text unavailable = 3
Not available in English = 2
Wrong setting = 1
Included
Eligible for inclusion (n = 62)
Studies included in qualitative synthesis (n = 72)
Additional records identified through hand searching (n = 10)

Figure 3.1 Study selection PRISMA flow chart for a review of the use of electronic sales records in population dietary surveillance – (June 2020)

Table 3.2 Summary of included paper characteristics

	Number of papers (%)
Country	
USA	33 (46)
Australia	8 (11)
New Zealand	6 (8)
Denmark	4 (6)
Finland	4 (6)
South Africa	4 (6)
UK	3 (4)
France	2 (3)
Italy	2 (3)
Netherlands	2 (3)
Barbados	1 (1)
Belgium	1 (1)
Canada	1 (1)
Switzerland	1 (1)
Year of publication	
1996 – 2000	11
2001 – 2004	1
2005 – 2008	2
2009 – 2012	9
2013 – 2016	24
2017 – 2020	25
Study design	
Policy evaluation	12 (17)
In-store choice architecture	16 (22)
Financial intervention	17 (24)
Feasibility	3 (4)
Dietary surveillance	16 (22)
Comparison with intake	2 (3)
Community intervention	6 (8)
Data aggregation level	
Country/Area	2 (3)
Store	25 (35)
Customer	42 (58)
Transaction	3 (4)
Socioeconomic status	
High	6 (8)
Mixed	8 (11)
Low	21 (29)
Not reported	37 (51)
Nutrient data source	

National FCDB	4 (6)
Commercial FCDB	1 (1)
Retailer (Back of Pack)	4 (6)
Combined	10 (14)
None	53 (74)
Duration of transaction data (months)	
0-12	46
13-24	15
25-36	6
37+	5

**FCDB = Food Composition Database*

3.3.3 Risk of bias

Of the 72 papers included in this review, the majority (42, 58%) were assessed as being of ‘Fair’ quality (Supplementary Table S5), this reflects the observational nature of many study designs. Ten papers (14%) in this review received a quality rating of ‘Good’, and the remaining 20 papers (28%) were rated ‘Poor’ in terms of quality of study design and reporting. Dominant risks of bias across studies were poorly defined study populations, lack of justification of sample size, and the reporting of participation and follow up rates.

3.3.4 Study aims

The majority of papers used transaction data to evaluate the success of dietary interventions (n = 40, 56%), including: financial incentives or penalties (n = 18, 25%), community behavioural change interventions (n = 7, 9%), or environmental nudges such as changes to the in-store architecture (n = 15, 21%) (Table 3.2). Twelve papers (17%) evaluated national or regional policies and 16 (22%) used observational designs for dietary surveillance (Table 3.2). Just two studies (3%) directly compared electronic transaction data with measures of dietary intake, and three investigated the methodological feasibility of using supermarket sales for dietary research by exploring methods for linkage with nutritional data sources (Table 3.2).

Intervention studies were typically short-term. Consequently, the majority of studies used no more than 12 months of transaction data (Table 3.2). A small number of studies collected transaction data over several years, with a maximum duration of eight years, these were typically policy evaluations or longitudinal dietary surveillance.

3.3.4.1 Evaluating intervention effectiveness

Transaction data provided evidence for success of in-store choice architecture interventions (Freedman and Connors, 2010; Payne et al., 2015; Vandenbroele et al., 2018; Van Gestel et al., 2018; Kroese et al., 2016), financial interventions (Ball et al., 2015; Brimblecombe et al., 2017; Le et al., 2016; Stead et al., 2017; Blakely et al., 2011; Franckle et al., 2018), and community interventions (Reger et al., 2000; Reger et al., 1999; Reger et al., 1998; Hobin et al., 2017; Dunt et al., 1999). By capturing all food purchases, transaction data revealed variation of intervention effectiveness by food category (Guan et al., 2018), with staple foods more resistant to change (Surkan et al., 2016; Dunt et al., 1999). Moreover, mode of intervention delivery is likely to influence effectiveness. For example, while online shopping shows promise for customisation of the shopping experience, through nudge-style interventions based on previous purchases (Moran et al., 2019), low-income customers possess a reduced tendency to shop online (Martinez et al., 2018). Thus, online interventions could widen societal inequalities. At the individual level, intervention effectiveness may be greater than purchase estimates suggest, as the size of individual dietary changes may be attenuated by household-level purchases (Reger et al., 2000).

3.3.4.2 Dietary surveillance

Electronic point of sale (EPOS) systems generate high-volume transaction data with a fine temporal granularity. Continuous transaction data revealed how dietary patterns change over time; including monthly trends in relation to payment in low income groups (Franckle et al., 2019), seasonally (Gamburzew et al., 2016; Schwartz et al., 2014; Franckle et al., 2019), and longitudinally (Walmsley et al., 2018). Thus, transaction records were used retrospectively for natural experiments in policy evaluations and provided commercial insights, including market trends (Frazao and Allshouse, 1996) and price elasticities (Jones, 1997; Revoredo-Giha et al., 2009; Guan et al., 2018). However, the degree of insight depends upon the level of data aggregation, both geographically and at the product-level.

Two studies (3%) aggregated supermarket purchase data to the country-level, for observation of national market trends (Frazao and Allshouse, 1996) and policy evaluation (Alvarado et al., 2019). Seven papers (10%) aggregated purchases to the area level (city or region), to understand the effectiveness of policies (Silver et al., 2017), community interventions (Reger et al., 1998; Reger et al., 2000; Reger et al., 1999; Dunt et al., 1999), and surveillance of regional dietary variations (Revoredo-Giha et al., 2009; Närhinen et al., 1999). No studies explored diet at the neighbourhood level, nor used geographic mapping techniques. Twenty-five papers (35%) used store-level purchases, to evaluate community interventions or policies, which employed cluster randomisation (Brimblecombe et al., 2017; Toft et al., 2017; Reger et al., 1998; Reger et al., 2000; Reger et al., 1999) and quasi-experimental designs (Silver et al., 2017; Alvarado et al., 2019; Brunello et al., 2014; Brunello et al., 2012; Mathios, 1998; Mathios, 2000; Balasubramanian and Cole, 2002; Ferguson et al., 2017).

Three papers (4%) disaggregated purchases to the transaction level. This increased data volume, permitting novel data-driven approaches to hypothesis generation, even without linkage to individual customers. For example, unsupervised machine learning (k-means clustering) revealed differences in dietary quality by type of alcohol purchased (Hansel et al., 2015; Johansen et al., 2006; Uusitalo et al., 2019). This suggests that, in a dietary patterns' context, alcohol type may be an important health consideration, perhaps as a marker for socio-economic status, in addition to total alcohol units.

In total, 42 papers (58%) used loyalty card records to link transactions at the customer level, via a unique customer identifier. Cohorts of loyalty card customers can be tracked over time, increasing confidence in observed temporal patterns and intervention effectiveness. Customer cohorts enable understanding of behavioural mechanisms, and reveal within-population dietary differences and intervention responsiveness. For example, the link between socioeconomic factors, intervention effectiveness (Gamburzew et al., 2016) and dietary quality (Phipps et al., 2014), suggests that restricting price promotions for unhealthy products may be more powerful for obesity prevention than discounting healthy products (Phipps et al., 2014).

In general customer demographics were poorly described. Over half of papers failed to report the socioeconomic status (SES) of participants (Table 3.2). Of those 42 papers using loyalty card data, 16 (38%) did not report any demographic information for the customer sample, hindering assessment of generalisability. Demographic information was most commonly obtained from baseline surveys (n = 23, 32%), which enabled researchers to capture sensitive information, such as BMI (Sturm et al., 2016), education level (Ball, Kylie 2016; Ball et al., 2015), and income (Ball et al., 2015; Le et al., 2016), that would not be held by the retailer. Two papers (3%) obtained demographic information from the retailer's records (Uusitalo et al., 2019; Nevalainen et al., 2018). Retailer captured demographic records were limited to age, gender and residential postcode (Uusitalo et al., 2019; Nevalainen et al., 2018). A further study attempted to use supermarket collected customer demographic information (Gamburzew et al., 2016), but was unable to do so due to poor completion of the loyalty card sign-up form. Additionally, customers forgetting to use their loyalty cards (Ni Mhurchu et al., 2007), and self-selection (Andreyeva and Luedicke, 2015) were identified as problematic for the coverage and generalisability of loyalty card customer samples.

In the absence of customer demographic information, thirteen studies used area-level proxies, based on store location. For example, area geodemographics (geographic segmentation based on the characteristics of people residing there) (Revoredo-Giha et al., 2009) or census tract characteristics (Andreyeva and Luedicke, 2013; Andreyeva et al., 2012; Silver et al., 2017) were used to characterise the customer-base.

Other socioeconomic proxies included store type (regular or discount) (Brunello et al., 2012; Mork et al., 2017) and payment method; such that payments made with an electronic benefits transfer card identified low-income customers in receipt of US state benefits (Andreyeva et al., 2012; Martinez et al., 2018; Polacsek et al., 2018; Moran et al., 2019). Four studies (6%) used geocoded store locations to reveal spatial and demographic variation in dietary behaviours (Närhinen et al., 1999) and responses to policy interventions (Brunello et al., 2014; Brunello et al., 2012; Mathios, 1998). No studies explored spatial variations in diet based on customer residential address.

3.3.5 Dietary assessment

3.3.5.1 Representativeness of total household purchasing

Four studies (6%) used additional self-reported household purchase data. They suggest that among loyal customers, supermarkets may account for between 63 - 67% of total household food expenditure (Ni Mhurchu et al., 2010; Ni Mhurchu et al., 2007; Hauser et al., 2013). However, shopping habits even among the most loyal customers are highly variable, resulting in wide confidence intervals around these estimates (Ni Mhurchu et al., 2007; Hauser et al., 2013). Additionally, missing data arising from technical issues with electronic data capture (Surkan et al., 2016; Ferguson et al., 2017) and customers forgetting to use loyalty cards during shopping, further reduced total purchase coverage by as much as 15% (Ni Mhurchu et al., 2007).

One study (1%) compared purchase records with national expenditure surveys (Hamilton et al., 2007). They reported that purchase estimates of proportional spend on staple foods fell within 2% of national expenditure surveys (Hamilton et al., 2007). However, agreement was poorer for discretionary products like sweet foods and beverages (Hamilton et al., 2007), even after excluding categories from the Household Expenditure Survey which were not covered by supermarket purchases (takeaway and restaurant meals) (Hamilton et al., 2007). No studies quantified the statistical agreement between household food purchase estimates from supermarket transaction records and self-reported expenditure.

3.3.5.2 Representativeness of individual food consumption

Seventeen studies (24%) collected additional data on self-reported individual dietary intake. No studies in this review attempted to extrapolate absolute dietary estimates from household purchases to the individual level. Instead, dietary estimates from purchase records were represented proportionally, such as percentage contribution to total energy or expenditure (Ni Mhurchu et al., 2010; Eyles et al., 2010; Hamilton et al., 2007). Other studies presented outcomes in terms of binary dietary behaviour indicators, i.e. customer purchased the food item of interest, or did not (Närhinen et al., 1999), or diet-quality indices (Chidambaram et al., 2013; Taylor et al., 2015).

One study (1%) directly compared household purchase estimates with individual self-reported consumption, using Spearman correlation coefficients and paired t-tests (Eyles et al., 2010). However, statistical agreement was not formally assessed (Eyles et al., 2010). Another study compared nutrient availability in supermarket purchases with national dietary consumption surveys (Hamilton et al., 2007). Overall, they reported good comparability between adjusted dietary estimates from purchase records and self-reported intake (Hamilton et al., 2007; Eyles et al., 2010). Yet there is evidence for variability in agreement by food type (Hamilton et al., 2007; Närhinen et al., 1999; Radimer and Harvey, 1998) and by nutrient (Eyles et al., 2010). Agreement was highest for energy from saturated fat and total fat. For protein, sugar and sodium, purchase records under-reported compared with repeated 24-hour recalls (Eyles et al., 2010), suggesting that key food sources of these nutrients are more likely to be purchased elsewhere. Contrastingly for other macronutrients, estimates from purchase records were higher than self-report estimates (Eyles et al., 2010).

Comparison with national dietary intake surveys also revealed differences in agreement within the population, with a poorer association observed for children's diets (Hamilton et al., 2007). Having children in the house is likely to affect the types of food chosen. A positive relationship was observed between purchases of fresh produce and the number and age range of children, independent of household size (Phipps et al., 2013). Household composition is therefore likely to be an important influencer of food purchasing and how products are distributed among the household, but this cannot be gained from secondary purchase records.

3.3.6 Sources of nutrient data

Of the 72 included papers, 53 (74%) did not link transactions to any source of nutrient information (Table 3.2). Four papers (6%) used National Food Composition Databases (FCDBs) only, three used 'Back of Pack' (BOP) product label information, one used information in the product description, and one used a commercial FCDB. The most common approach was to combine multiple data sources (10 papers, 14%) (Table 3.2), creating a custom FCDB with which purchased food and beverage products could be matched.

The source of nutrient information influences the degree of error incorporated into dietary estimates at the nutrient level. National FCDBs are used to code dietary survey responses as they contain detailed nutrient information for commonly consumed generic foods. Yet, matching to transaction records results in reduced dimensionality from several thousand retail products to just a couple of thousand generic foods and a loss of product-specific detail

(Gamburzew et al., 2016; Brinkerhoff et al., 2011). Furthermore, FCDBs are restricted to the most commonly consumed foods and may therefore poorly represent ethnic foods (Tran et al., 2017). This introduces greater error into nutrient-level estimates for some population sub-groups. Despite these limitations, national FCDBs are readily available, enable comparison with national dietary surveys (Chidambaram et al., 2013) and adjustment for edible portion and specific gravity (Brimblecombe et al., 2017; Tran et al., 2017), improving the representation of products as eaten rather than as sold.

However, matching transaction data to FCDBs is challenging. Due to the large number and high turnover of retail products, there have been attempts to develop automated, scalable and repeatable FCDBs matching approaches. While near-perfect matches for standard food groups may be possible, in the absence of commonly used product identifiers, there are barriers to mapping to detailed nutrient content (Brinkerhoff et al., 2011; Brimblecombe et al., 2017). At the food item level, string- and fuzzy-matching algorithms may be hindered by retailer abbreviations (Brinkerhoff et al., 2011; Tran et al., 2017; Chidambaram et al., 2013). This may be overcome if a full product description can be identified from the unique product code (UPC) by web-scraping (Chidambaram et al., 2013). Nevertheless, in some circumstances, retailer short product descriptions can prove advantageous in minimising noise from excess information which reduces match accuracy (Tran et al., 2017).

Alternatively, nutrient data may be mapped at the category or sub-category level (Tran et al., 2017; Brinkerhoff et al., 2011) However, this is prone to mis-matching errors resulting from different categorisation approaches (Taylor et al., 2015; Brinkerhoff et al., 2011). FCDB categories are nutritionally-led, while retailer categories are based on product placement in store and are consequently nutritionally heterogeneous (Taylor et al., 2015; Brinkerhoff et al., 2011). For example a retailer ‘soft drinks’ category, including both full sugar and diet beverages, resulted in a mis-match of around 30% (Brinkerhoff et al., 2011).

Where BOP nutrient information is available from the retailer, automated linkage to the transaction record may be achieved via the UPC (Banerjee and Nayak, 2018; Eyles et al., 2010; Hamilton et al., 2007; Ni Mhurchu et al., 2010; Hobin et al., 2017; Chidambaram et al., 2013). This improves product-specific nutrient accuracy and coverage of the product portfolio. In turn, this enables between-brand comparison, and reflects changes in formulation over time (Jones, 1997; Johansen et al., 2006). However, the ever-evolving retail offer makes UPCs an unstable identifier (Brinkerhoff et al., 2011). Furthermore, lack of publicly available digitised UPC-level FCDBs was highlighted as a major barrier to linkage between transactions and their nutrient values (Brinkerhoff et al., 2011; Chidambaram et al., 2013). While commercial datasets are available (Hobin et al., 2017), cost and data sharing agreements restrict their use (Chidambaram et al., 2013) and their availability cannot be relied upon. Since their publication, two of the third-party data sources used by studies in this review are no longer available for use (Hamilton et al., 2007; Banerjee and Nayak, 2018). For these reasons, a combination of nutrient data sources was typically used by researchers, generating their own FCDB (Hamilton et al., 2007; Banerjee and Nayak, 2018; Silver et al., 2017; Andreyeva et al., 2012; Andreyeva et al., 2014; Gamburzew et al., 2016; Brinkerhoff et al., 2011; Brimblecombe et al., 2017; Tran et al., 2017; Chidambaram et al., 2013).

3.3.7 Outcomes

Nutrient-level analyses (Banerjee and Nayak, 2018; Hamilton et al., 2007; Eyles et al., 2010; Ni Mhurchu et al., 2010; Hobin et al., 2017) focused on energy and key BOP macro-nutrients. With the exception of sodium (N = 3) (Eyles et al., 2010; Brimblecombe et al., 2017; Banerjee and Nayak, 2018), no studies conducted micro-nutrient level analysis. Nutrient analyses were presented in absolute terms at the household level (Banerjee and Nayak, 2018), or more commonly were energy-adjusted, meaning that nutrient-specific dietary adequacy could not be assessed.

Due to challenges of data availability and linkage with nutrient data, most studies conducted analysis at the food category or sub-category level (Hansel et al., 2015; Payne et al., 2015; Taylor et al., 2015). As Brinkerhoff et al. (2011) describes, supermarket-derived categories may not be wholly meaningful from a nutritional perspective. Category-level purchases were measured in terms of relative or absolute unit sales (Surkan et al., 2016; Hobin et al., 2017; Vandenbroele et al., 2018), expenditure (Payne et al., 2015; Ferguson et al., 2017; Polacsek et al., 2018) or weight/volume/portions (Brunello et al., 2014; Brunello et al., 2012; Ball et al., 2015; Silver et al., 2017). Single food products (Vandenbroele et al., 2018) or broader categories (commonly fruit and vegetables (Ni Mhurchu et al., 2010; Walmsley et al., 2018; Brimblecombe et al., 2017) and soft drinks (Ferguson et al., 2017; Ball et al., 2015; Alvarado et al., 2019)) were used as outcomes for intervention and policy evaluations. As food purchase decisions are not independent of each other, this approach may miss unintended negative consequences such as substitution effects within other categories (Andreyeva and Luedicke, 2013; Andreyeva and

Luedicke, 2015). For this reason, Taylor et al. (2015) advocates a broader dietary pattern view to examine dietary quality.

The study of dietary patterns involves classifying customers into groups based on their purchase habits. Groups may be defined a-priori, based on the purchase of some product of interest, in a deterministic approach. For example, Johansen et al, (2006) used a dichotomous approach based on whether items were purchased or not, to classify customers as wine-buyers, beer-buyers, mixed or non-alcohol purchasers. Instead, groups may be defined based on the dietary quality of products purchased. Products may be classified on evidence for diet-disease relationships (Hansel et al., 2015), professional opinion(Surkan et al., 2016) or using custom or established Nutrient Profile Models (NPMs) (Phippset al., 2014; Gamburzew et al., 2016; Taylor et al., 2015; Hobin et al., 2017; Chidambaram et al., 2013; Franckle et al., 2018). However, in many cases, classification criteria were not transparently described for reproducibility. Established NPMs use pre-defined criteria, making them stable metrics for dietary surveillance; for example, in assessing compliance with dietary guidelines. Classification of products shows that shoppers prioritise purchases of ‘unhealthy’ food products over ‘healthy’ foods (Hansel et al., 2015; Taylor et al., 2015). The majority of expenditure was on discretionary foods (34.8%), followed by meat and meat alternatives (17.0%), with the least spent on vegetables and dairy products (Taylor et al., 2015). Vandenbroele et al, (2018) advocated retailers shift from product-focused thinking to a whole basket approach. Focusing on overall purchase dietary quality will enable retailers to implement choice architecture strategies which maximise health as well as profits.

Alternatively, dietary patterns may be explored non-deterministically through unsupervised machine learning algorithms. Hansel et al. (2015) used K-means clustering to classify customers according to their alcohol purchase habits. Not only does this approach account for frequency and quantity, it revealed greater dietary nuance between alcohol purchasing groups, such as the specific dietary habits of purchasers of aniseed-based beverages and Bordeaux wines. Studies observed a relationship between purchases of beer and the less healthy traditional-type diet and between purchases of wine and the healthier Mediterranean-type diet (Hansel et al., 2015; Johansen et al., 2006), highlighting the utility of dietary patterns for describing dietary quality, although they cannot quantify it.

3.4 Discussion

The 2017 Review of Nutrition and Human Health Research (MRC, 2017) describes a field in crisis. The review highlights the limitations of self-reported dietary intake methods which, it is argued, contribute to a perceived lack of rigour in nutrition research (MRC, 2017). The inability to accurately measure diet has damaged confidence in nutrition research findings. Consequently, there has been little progress towards improvements in population diet, despite substantial efforts from interventions and policy (Theis and White, 2021).

There is no gold standard method of dietary assessment which can answer all diet-related questions. The breadth of questions posed by the field of nutrition research therefore require a suite of innovative methods to supplement existing approaches. This necessitates the harnessing of technology and secondary data sources where they are available. Just as biomarkers complement surveys with an objective measure of nutrients within the body, supermarket transaction records provide a complementary objective measure of food purchases.

This review found that supermarket electronic purchase records can be useful for longitudinal dietary surveillance (Radimer and Harvey, 1998) in high and middle-income populations where supermarket shopping is prevalent and represents the majority of household expenditure (Ni Mhurchu et al., 2010; Ni Mhurchu et al., 2007; Hauser et al., 2013). Transaction data has a number of strengths. Large data volumes enable data-driven exploration of dietary patterns (Johansen et al., 2006; Morris et al., 2020; Hansel et al., 2015; Uusitalo et al., 2019) to better understand food purchase behaviours and identify intervention target groups. Furthermore, continuous data collection permits observation and control for day to day (Närhinen et al., 1998), week by week (Franckle et al., 2019), and seasonal variation in dietary choices (Sturm et al., 2016) which cannot be revealed in such detail by cross-sectional dietary surveys.

Large customer samples and passive data collection may improve representation of hard-to-reach groups. This was demonstrated by good diffusion across income groups within a single-retailer sample in the UK (Jenneson, et al., 2020; Jenneson, et al., 2021), despite being unrepresentative of the general population overall (Green et al., 2020). Similarity to regional dietary estimates from survey data (Närhinen et al., 1999), highlights the utility of electronic transaction records for within country ecological study of diet. To date, much research into spatial variation in diet has focused on the food environment (Wilkins et al., 2019), such as accessibility to supermarkets (Aggarwal et al., 2014) or fast-food outlets (Fraser et al., 2010) rather than actual behaviours. While spatial patterns may be observed in dietary survey data (Morris et al., 2016), large sample sizes are required to reduce the risk of ecological fallacy and as such, aggregation areas tend to be large. Using store location and, where available, customer area of residence (as geo-coded reference points) the scale of electronic supermarket purchase data enables small-area spatial analysis,(Jenneson et al., 2021; Jenneson et al., 2020; Clark et al., 2021) provided this is permitted by data usage agreements, given the proprietary nature of retailer data, and that appropriate information management systems are in place.

Limitations of electronic supermarket transaction data are; partial coverage of total food purchased or otherwise obtained, unknown distribution of food within households, and inability to account for food wasted, or food consumed by visitors (Greenwood et al., 2006). As such, household-level purchase data does not directly measure individual dietary intake (Eyles et al., 2010). Studies in this review suggest, at best, moderate agreement between household purchase and individual intake estimates (Eyles et al., 2010; Hamilton et al., 2007). Given these limitations, there is a need for validation against existing methods to better understand the utility of supermarket transaction records for monitoring dietary behaviours. Triangulation with other dietary assessment methods may reveal additional insights and enable generation of adjustment factors for improved consumption estimates.

Statistical agreement (Bland, 1986) between electronic purchase records and self-reported methods was not formally assessed by studies in this review. However, observed correlations (Eyles et al., 2010; Hamilton et al., 2007) support its ability to capture the majority of the diet. This adds weight to earlier work which found good agreement between estimates of fat and energy from paper-based cash register receipts and self-reported 4-day food diaries (Ransley et al., 2001). Just how much purchase data is required to represent habitual diet warrants further exploration, but evidence from this review suggests that around 7 days of transaction records may be enough to represent usual diet (Närhinen et al., 1998), at least for perishable high-turnover products.

As no studies in this review attempted to adjust household purchases to the individual level, it is unclear how well household purchases represent the diet of individuals within a household. Modelling individual diet from household purchases would require a number of assumptions, which necessitates further study (Greenwood et al., 2006). To do so, additional survey information (Ransley et al., 2001) about household composition and within-household food distribution would be needed to adjust for person-specific measurement error (Greenwood et al., 2006). Alternatively, modelling techniques, such as microsimulation (Smith et al., 2006) and other mathematical approaches may offer a means to estimate diet at the individual level. Transaction data can contribute to refinement of modelling parameters, for example, understanding the impact of age of children in the household on fruit and vegetable purchase quantities (Phipps et al., 2013).

There is increasing recognition of the importance of engaging with the food industry to translate research insights into action. Effective research-industry partnerships are therefore vital, as explained by the guidance framework proposed by Birkin et al.(2019). While the studies in this review did not explicitly discuss the challenges associated with partnership building, it is a key consideration for researchers wishing to harness the potential of supermarket transaction records. That said, challenges relating to the way data are shared can contribute to the sources of bias observed by this review; a lack of information about the study participants, lack of transparency in recruitment, and inability to control for customer demographic characteristics, which might act as confounders for dietary behaviours.

New approaches to transparency and customer consent are therefore warranted to enable greater utility of customer-level data. Efforts are needed to overcome issues of poor data quality (Ni Mhurchu et al., 2007; Gamburzew et al., 2016), restricted information (Uusitalo et al., 2019; Nevalainen et al., 2018) and assessment of customer sample bias. Innovations such as the Danish Data for Good Foundation's platform, which enables bespoke customer informed consent and triangulation of public and private-sector data, could offer a potential solution (Data For Good Foundation., 2021).

Another obstacle for the future of the method is the lack of centralised and up to date product-level food composition databases which may be linked to automatically (Greenwood et al., 2006). Studies in this review reported the need to create new bespoke FCDBs to facilitate linkage with nutrition information. This requires a substantial amount of up-front resource which limits time to generate interesting research insights. While commercial FCBDs exist, cost and data sharing agreements can be a barrier. Furthermore, their coverage is typically limited only to those nutrients required to be reported by local BOP labelling regulations, which contributes to a lack of utility for micronutrient monitoring, and differences in nutrient coverage between countries. In contrast, national food tables are freely available and cover a wider range of nutrients, but for fewer and more generic foods.

Solutions could include the linkage of product data to close-matching generic foods in national FCDBs which contain detailed micronutrient compositions, as performed by the dietary assessment app myfood24 (Carter, 2016). Yet, ensuring these stay up to date remains a challenge. Innovations such as FoodDB (Scarborough, 2021) harness web-scraping to provide regularly updated BOP nutrition composition information for products on the market. It is also possible that 3D barcode advances, which permit greater data capture, may further improve product-level FCDBs in the future through the inclusion of micronutrient information and supply chain data, such as origin and sustainability metrics. Viable country-specific FCDB solutions are therefore vital to enable nutrient and brand-level research insights from supermarket transaction data, which this review found to be lacking. In addition, there is a role for bodies such as the FAOs International Network of Food Data Systems (INFOODS) (FAO, 2021) to develop global standards around the reporting and exchange of product nutrient data to promote consistency and facilitate across-country comparison.

3.4.1 Future research priorities

This review highlights five priority areas for future research into the use of supermarket sales data for population dietary surveillance: 1) validation against established self-report methods and nutritional biomarkers; 2) extrapolation of household purchases to the individual level; 3) triangulation with other data sources; 4) exploration of spatial dietary patterns; 5) development of suitable nutrient datasets for linkage.

3.5 Conclusion

This review suggests that electronic purchase records have broad applicability for dietary surveillance, policy evaluation and intervention research studies in high- and middle-income countries. The scale, temporality and geocoded nature of electronic purchase records are notable advantages. However, there is a need for further methodological assessment of utility; validation against self-reported dietary intake measures and nutritional biomarkers; required data volumes; extrapolation to the individual level; exploration of spatial dietary patterns; and assessment of generalisability. The potential for automated dietary coding is currently hindered by the availability of regularly updated open product data. Web-scraping methods may address this need. However, this limits coverage to key back of pack nutrients, which excludes micronutrients (with the exception of sodium). Product data alone accounts only for dietary availability; linkage with sales data is crucial for behavioural research.

3.6 Acknowledgements

3.6.1 Funding

Funded by an Economic and Social Research Council Grant (ES/R501062/1) awarded as part of the Centre for Doctoral Training Programme in Data Analytics and Society. The funders played no role in the conduct of the systematic review.

3.6.2 Declaration of interest

Victoria Jenneson and Michelle Morris declare their work in partnership with UK national retailers.

3.7 Supplememtary materials

Table S1 – PRISMA Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	2 - 4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	4
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	4
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	4, (Appendix 3)
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	4
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix 2

Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	5
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	Appendix 4
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	5 – 6, Appendix 6
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	N/A (narrative synthesis)
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	N/A
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	7
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/A
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	6 (Figure 1)
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Appendix 5
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Appendix 6
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	7-15

Synthesis of results	21	Present the main results of the review. If meta-analyses are done, include for each, confidence intervals and measures of consistency.	Appendix 5
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	7
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	N/A
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	16 – 23
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	16 – 23
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	23
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	24

Reported page numbers refer to pages in article as submitted

Table S2 – Example search strategy (MEDLINE; OVID interface, 1996 onwards

#	Search Term
1	diet$.mp or DIET/ or “DIET, FOOD AND NUTRITION”/
2	diet records.mp or Diet Records/
3	energy intake.mp or Energy Intake/
4	food.mp
5	diet quality.mp
6	Nutrition Assessment/ or dietary assessment.mp
7	food supply.mp or Food Supply/
8	(food adj purchas$).mp
9	(diet surveys or nutrition surveys).mp
10	nutrition monitoring.mp
11	((food or diet$) adj habit$).mp
12	or/1-11
13	Commerce/ or supermarket$.mp
14	grocery store$.mp
15	shop$.mp
16	food industry.mp or Food Industry/
17	or/13-16
18	sale$.mp
19	purchas$.mp
20	(scan$ adj data).mp
21	receipt$.mp
22	(loyalty adj card).mp
23	or/18-22
24	and/12, 17, 23
25	Limit 24 to “all adult (19 plus years)”

Table S3 – Data extraction form

Study intention	Study aims What was the study designed to assess? Are the aims clearly stated?
	Describe location & setting. Might this target/exclude certain groups?
	Start and end date of study
	Total study duration
Methods	Method of participant recruitment (does this differ by setting?)

	Inclusion/exclusion criteria for participation
	Representativeness of sample: are participants likely to be representative of the target population?
	Total number of (intervention) groups
	Sample size (for each group)
	Was randomisation used? If so, what unit (individuals or cluster/groups)?
	Unit of analysis? Aggregation level, geographic unit. (Where applicable, was this the same as the randomisation unit?)
	Describe intervention / control conditions where relevant (setting, theory, delivery, timing etc)
	Statistical analysis methods used. Were these appropriate?
	If secondary analysis, what was the original data purpose & context? Could this introduce bias?
Results	What percentage of participants agreed to participate?
	Were there any significant baseline imbalances between groups?
	What percentage of participants completed the study?
	Describe participant characteristics (for each group)
	Definition of outcome(s) including units of measurement and unit of aggregation if relevant
	How were outcomes measured?
	Time points measured
	Results
Other relevant information	Potential for author conflict? Would one outcome benefit authors/data collectors?
	Author's key conclusions
	Comments from review authors

Table S4 – Summary of included studies

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
Policy evaluations											
Nutrition Labelling and Education Act (NLEA)	Mathios (1998)	Cross-sectional	Impact of NLEA on type of cooking oil purchased	USA, New York State 20 stores	Customer number unknown Loyalty card demographics used for sampling to ensure breadth – based on educational attainment.	2 years, collected every 4 months (Oct 1992 – Oct 1994)	Store-level	Product nutrition labels	Market share-weight (units sold) of fat in oils (saturated, mono-unsaturated, poly-unsaturated)	Econometric model, regression analysis	Saturated fat increased all stores. Mono-unsaturated declined 17/20 stores. Least educated increased saturated fat and most educated increased mono-unsaturated.
	Mathios (2000)	Quasi-experimental	Impact of NLEA on sales of salad dressings						Market share (units sold)/ week Fat (g/serve) % products with voluntary nutrition label	Correlation between per serving fat and calories Econometric model, regression	Correlation 97% Sales of unlabelled products higher for less educated supermarkets Greater reduction in market share for products highest in fat
	Balasubramanian and Cole (2002)	Longitudinal	Impact of NLEA on sales of products with specific	USA, several stores, major grocery	Number & customer demographics unknown	7 years 8 months, weekly sales	Store-level weekly scanner data	Claim in category or product description Y/N	Category share with 10 week moving average	Regression analysis	Increased sensitivity to negative nutrient claims, purchases of positive nutrient

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
			nutrition descriptors e.g. low fat	chain in large city							products declined or stable
European School Fruit Program	Brunello et al. (2012)	Controlled before and after	Effect of EU School Fruit Program on sales of unhealthy snacks	Italy 44 stores, 2 retailers (one discount, one regular)	15 treatment stores (within 500m of treated school) Year 1 n=100 29 control stores, Year 1 n=479, Year 2 n= 405	2 years (Jan 2009 – Sept 2011) 1-year pre- & 1-year post	Aggregated store-level sales data	N/A	Mean daily store sales of unhealthy sweet and salty snacks (units, kg)	Difference in differences Regression	Treated stores 4.6% reduction in snack vs control (not significant) Significant reduction in high income areas (-12%), regular stores (-13%) and branded products (-13%). No effect in low income areas and discount stores.
	Brunello et al. (2014)										
Berkeley sugar tax	Silver et al. (2017)	Interrupted time series	Impact of Berkeley sugar sweetened beverage (SSB) tax (1 cent/ounce) on sales, price & intake	USA 26 stores Berkeley California; 3 Berkeley intervention stores, 6 control stores outside Berkeley	N = 957 Adults living in Berkeley. Affluent city with high education and low baseline SSB intake	3 years Pre- and post-taxation	Daily point of sale data Store price surveys 2 repeated 24-hr dietary recall telephone surveys	Nutrition data from product website, nutrition facts panel from Mintel, USDA database	Changes in inflation-adjusted prices (cents/ounce) for taxed SSBs, sales (ounces), customer spend/ transaction, intake (g/day and kcal/day)	Difference in difference OLS regression volume & revenue per transaction for Berkeley vs non-Berkeley stores	Taxed sales fell by 9.6%, untaxed rose by 3.5%. No change in customer spending or store revenue. Mean intake (g) reduced by 19.8% & calories from SSBs fell by 13.3% (both non-significant)

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
Special Supplemental Nutrition Program for Women, Infants, and Children (WIC)	Andreyeva et al. (2012)	Cross-sectional	Compare non-alcoholic beverage purchases for WIC vs SNAP benefit recipients	USA Large supermarket chain, several New England states	39,172 loyalty card holders Low income young families eligible for federal food & nutrition assistance	6 months (January – June 2011)	Loyalty card scanner data	Gladson's Nutrition Database + internet searches	Refreshment beverage purchases/hh/month	Generalised linear regression from Poisson family with logarithmic link function	64% matched to nutrient data SNAP household purchased more (689 oz) than WIC (352 oz) and more SSB (58% vs 58% respectively) SNAP paid for 72% of SSBs, ~ $1.7 – $2.1 billion/year
	Andreyeva et al. (2013)	Natural experiment	Effect of reduced juice allowance for the Women, Infants and Children (WIC) programme	USA >60 stores from one chain in Connecticut & Massachusetts	2137 households Loyalty card holders Low income young families eligible for WIC	20 months (January 2009 – September 2010)	Loyalty card scanner data	N/A	100% juice purchases (floz/hh/month) by payment type (%)	Generalised linear regression from Poisson family with logarithmic link function	Total juice declined 23.5% (21.4% - 25.4%) Reduction in 100% juice & WIC proportion. Small increase in non-WIC juice, fruit drinks, & non-carbonated. 12% (8.1% - 15%) decline in soft drinks
	Andreyeva and Luedicke (2013)		Impact of including whole-grain products in WIC on purchases					Gladson's Nutrition Database, internet searches, My Pyramid	Bread (whole grain 100%, 51-99%, 1-50%/white) & rice (brown or		100% whole grain bread share tripled; 8% - 24%. White bread fell; 58% - 50%. Overall bread

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
			of bread & rice					Equivalents Database	white) purchases/hh/month		stable. Decline in non-WIC Brown rice share rose (0.3 Oz to 2.4 Oz), rise in white & total rice
	Andreyeva et al. (2014)		Impact of reduced WIC milk & cheese allowance & disallowance of whole milk over 23 months					N/A	milk (floz) & cheese (oz) purchases/hh/month, share of whole milk, saturated fat from milk & cheese (g)		13% reduction in total milk and 20% reduction in WIC-milk purchases. Significant reduction in whole milk share and 40% reduction in WIC-eligible cheese purchases
	Andreyeva and Luedicke (2015)		Impact of WIC fruit & veg vouchers on fruit & veg purchases					N/A	fruit & veg purchases/hh/month (weight, cup equivalents & expenditure)		Fruit & veg increased significantly (+17.5% & +28.6% respectively) P<0.001
Barbados Sugar Tax	Alvarado et al. (2019)	Interrupted Time Series	Impact of 10% added value tax on Sugar Sweetened Beverages (SSBs)	Barbados	Barbados shoppers – demographics unknown	3 years 10 months	Country-level sales from one grocery chain	N/A	Weekly sales volume (mL) per capita SSBs and non-SSBs	Interrupted time series, linear regression	SSB sales decreased 4.3% (-4.9, -3.6%) Non-SSB sales increased 5.2% (4.5, 5.9%)
Financial interventions											

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
Supermarket Healthy Options Project (SHOP)	Ni Mhurchu et al. (2007)	Pilot RCT	Promote healthier purchases: culturally appropriate nutrition education & 12.5% price discount	New Zealand 5 Shop 'N Go stores	95 hhs Age (µ) 40yrs, 72% female, 7% Maori, 2% Pacific, 91% European / other	6 months (12 weeks baseline, 12 weeks intervention)	Self-scan transactions	N/A	Total hh food expenditure & fruit and veg purchases	Participant descriptive statistics Analysis of shopping diaries	Poor enrolment by minority ethnic groups. Supermarkets = 66% total expenditure (51% captured by Shop 'N Go system, 33% at other retailers)
	Ni Mhurchu et al. (2010)	RCT	Effect of 12.5% price discount & tailored nutrition education on food & nutrient purchases	New Zealand 8 Shop 'N Go stores	1,104 households, Age (µ) 44yrs, 86% female, 22% Maori, 9% Pacific, 68% European/ other, 52% low income, 51% low qualification	12 months + 12 weeks baseline (24 weeks intervention, 24 weeks follow up)		Supermarket Food & Nutrition Database (SFND); Manufactured Food Database, brand websites, back of pack & NZ food tables	% hh food energy from saturated fat (other macro-nutrients secondary) purchases of 'healthier' food (kg/hh/wk),	Repeated-measures mixed-model (difference from baseline) regression Intention to Treat analysis	Increased healthy food, & fruit & veg purchases (+11% & +15% respectively) 6 months vs baseline, no difference in saturated fat or other macronutrients
	Blakely et al. (2011)	RCT								Sensitivity analysis by SES (ANCOVA)	Effect varied by ethnicity (non-significant); Maori -0.15kg/wk (CI -1.10, 0.8), Pacific +1.20kg/ wk (CI 0.06, 2.23), European/other +1.02kg/wk (CI 0.60, 1.43)

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
Supermarket Healthy Eating for Life (SHELf)	Ball, K. et al. (2015)	RCT	Cost-effectiveness of tailored skill-building & price reduction to promote purchase & consumption of healthy foods & beverages among high- & low-SES women	Australia, Coles stores 2 target stores	N = 574 female loyalty card holders, Age (µ) 43.7yrs, 44.4% low SES catchment, 50.1% tertiary education, 28.6% born outside Australia	3-month intervention, 6 months follow up, 3-month retrospective baseline data	Loyalty card transactions FFQ & self-reported soft drink portions, Questionnaire	N/A	Purchase & consumption / hh/wk of fruit & veg (g), sugar-sweetened & low-calorie soft drinks, water (serves, ml), self-efficacy & perceived affordability	Generalised Estimating Equations Mediation analyses (MacKinnon method)	Increased fruit purchases at 3 months +35% (2.4 serves/wk) & veg +15% (3.1 serves/wk) vs control. Self-reported fruit consumption increased (+2.43 serves/wk). No increase in diet beverages or water. No difference by income or education
	Le et al. (2016)	RCT							ICER (incremental cost-effectiveness ratio) A$/ additional serve	Bootstrapping with 1000 resamples. Cost-effectiveness plane.	Price Reduction: ICER = $2.3 per extra serve veg/wk , $3.0 per extra serve fruit/wk Combined: ICER = $11.6 per increased fruit serve/wk
NYC supermarket discount	Geliebter et al. (2013)	RCT	Effect of 50% price discount on purchase & intake of low-energy density	USA 2 stores Manhattan, New York (~1 mile apart)	N = 47 loyalty card holders, 70% female, BMI (µ) 30.2, Age (µ) 37.5, 56% Caucasian,	16 weeks (4 weeks baseline, 8 weeks intervention, 4	Loyalty card transactions, continuous over 16-weeks	N/A	Gross hh expenditure ($/wk), intake (g, kcal & servings of fruit & veg	ANOVA with repeated measures (95% significance level)	Fruit & veg purchases in discount group increased 3x vs control, intake +1.5 serves. Purchases & intake significantly

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
			fruit & veg, bottled water & diet sodas, & body weight		19% African American, 13% Hispanic 28 Intervention, 19 Control	weeks follow up)	5 repeated 24-hr recalls (4 weeks apart)		/day, 1 serve = 80g) Body weight (kg), BMI, Body fat %		correlated (r = 0.62). No difference in beverage purchase or intake
	Bernales-Korins et al. (2017)	RCT	Effect of 50% discount on purchase & intake of fruit & veg		N = 45 loyalty card holders		As above plus psychosocial measures (determinant of intake)		As above, plus self-efficacy, stages of change & perceived barriers	As above, plus structural equation modelling	Discount increased self-efficacy & stages of change but no change in perceived barriers.
SHOP@RIC (Stores Healthy Options at Remote Indigenous Communities)	Brimblecombe et al. (2017)	RCT	Effect of 20% price discount on food & drink purchases with & without consumer education	Australia 20 remote indigenous communities with single store (2 retailers, ALPA & OBS)	Combined population ~8,515 people 10 stores discount, 10 stores discount + education	2.5 years 19 weeks baseline, 24 weeks intervention, 24 weeks post-intervention	Sore-level weekly sales data	Food Standards Australia & New Zealand Australian Food, Supplement and Nutrient Database 2011-13	Primary = per capita daily weight (g) fruit & veg purchased Secondary = beverages, healthy/ unhealthy foods (g/day)	Mean difference (name of test not given)	Increased fruit & veg during (+12g/capita/ day) & after intervention (+18g/ capita/day). More effective for fruit. Additional benefits for veg with education. Increase in total beverage, total unhealthy products, sodium & energy too

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
Outback Stores	Ferguson et al. (2017)	Natural experiment	Evaluate price strategies; reduced grocery mark-up, fruit & veg scales, fruit & veg sold at landed cost, diet soft drink discount	Stores in 18 remote Aboriginal communities, Central & Northern Australia	18 stores, 54 interview participants, 78% aboriginal, 89% over 35 years, 48% male	18 months (July 2009 – December 2010)	Store-level monthly sales data	N/A	Change in grocery sales ($) ratio of total sales, fruit & veg sales, soft drink sales	Mixed effects model with random effect intercept, adjusted for correlation of monthly sales with same store Autoregressive model, controlling for season	No impact on sales/turnover of grocery, fruit & veg or soft drinks.
Buywell trial	Stead et al. (2017)	RCT	Assess impact & feasibility of targeted price promotion & healthy eating advice on targeted healthy foods	Scotland, UK Low income areas	N = 53,363 loyalty card customers who purchase unhealthy products, 31 – 65 years 37,034 intervention, 16,333 control	6 months 2 months baseline, 1- month intervention, 3 months follow-up	Loyalty card EPOS transactions	FSA traffic light scheme & nutrient profiling used for population sampling only	No. & % customers purchasing targeted healthy products, Product switching	Chi-squared	Significant increase in proportion purchasing 4/5 targeted products. No significant increase for fruit & veg 8% customers switched to lower fat milk during intervention. Effects not sustained.
Sylacauga Aliance for Family Enhancement (SAFE)	Banerjee and Nayak (2018)	RCT	Effectiveness of targeted education & price discount on	Alabama, USA 2 local stores	N = 100 low income families 83% female, 31% Caucasian,	1 week	Scan data linked with store card, issued specifically for study	Food-A-Pedia or product nutritional label	Change in kcal/hh, sodium, added sugar, saturated fat & fibre (alcohol &	Linear regression	Education & combined significantly reduced total calories (235 & 280kcal) respectively, vs

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
			purchases of healthy food		65% African American, Age (µ) 39 yrs, 70% unemployed 25 Education, 25 Coupon, 25 Combined, 25 Control		with $40 credit		non-food items excluded)		control. Coupon reduced by 97kcal relative to control (non-significant)
Healthy Food Program	Sturm et al. (2013)	Case-control	Effect of 10% & 25% price reduction of healthy food on household shopping behaviours	South Africa >400 stores (single retailer)	Members of private health insurance Healthyfood programme; No rebate: N ~ 67,794, 10% rebate: N ~ 33,558, 25% rebate: N ~ 68,133 No demographics	3 years (2009 – 2012)	Purchases from eligible supermarkets using specific credit card, for linkage with health record & rebate Health Risk Assessment Survey (HRA)	N/A	Ratio of healthy, fruit & veg, neutral, less-desirable, to total spend /hh/month Intake servings fruit & veg, & whole grain, salt, sweet foods, processed meat & fast food. Self-reported weight & height	Household fixed-effects model & case-control difference-in-differences Sensitivity analysis; proximity to eligible stores & customer loyalty	Negligible bias by payment type or strategic shopping 10% rebate: +6% healthy, +5.7% fruit & veg, -5.6% less-desirable 25% rebate: +9.3% healthy, +8.5% fruit & veg, -7.2% less-desirable
	An (2014)									Descriptive statistics: two-sample t-test with unequal variance	Members ate more fruit & veg (+0.7 serves), 8% more likely to meet wholegrain guidelines, 2% less likely to eat high sugar, fried foods (-9%), salt (-2%), processed

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
											meat (-7%) & fast food (-8%)
	Schwartz et al. (2014)	RCT	Effect of voluntary self-control financial commitment (forfeit 25% healthy food rebate) on healthy food purchasing	South Africa >400 stores	N = 4,073 households, members of private health insurance programme, no demographic information, 62% completed	12 months (6 months intervention + 6 months baseline)	Transactions at FlyBuys supermarket using Discover-Health visa credit card		Household purchases (%) of healthy/ neutral/ unhealthy foods (units & expenditure)	Intention to treat analysis, random-effects linear regression	36% hhs accepted pre-commitment Pre-commitment hhs increased healthy purchases by 3.5%. No change among control or those who declined
Healthy Incentives Pilot (HIP)	Bartlett (2014)	RCT	Effect of financial incentives (30% point of sale rebate) for benefit recipients, on consumption of fruits, veg & other healthy foods	USA, Hampden County, Massachusetts 130 intervention stores	55,095 SNAP households; 7,500 intervention group (HIP), 47,595 control group (non-HIP) Mean age 43, 73% female heads of household, ~50% Hispanic	1 year (2011 – 2012)	Electronic Benefit Transfer Card (EBT) Self-reported consumption & spend; telephone 24-hour recalls	N/A	Consumption (cups/day) on targeted fruit & veg (TFV) EBT ($) on TFV Self-reported ($) on total fruit & veg	Regression-adjusted differences between HIP and non-HIP groups	HIP consumed 0.25 cups more TFV/day (+26%), spend $6.15 more/month, & +$1.19 more EBT spend on TFV than non-HIP

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
Reducing SSB consumption in Latino community	Franckle et al. (2018)	RCT	Effect of financial incentive and traffic light labelling scheme on reducing purchases of sugar-sweetened beverages	USA, Boston, Massachusetts 1 store	148 households with children under 18yrs, shopping in low-income community Intervention: 100% female, 34% over 40 years Control: 97% female, 34% over 40 years	2 months baseline, 5 months post-intervention	Study-specific loyalty card Exit interview, self-reported consumption	N/A	% customers purchasing beverages labelled with red traffic light (>12g sugar/12oz serve) each month Binary outcome ≥1 serve or none	Logistic regression	Difference in purchases of red-labelled beverages between groups (p=0.002). Intervention group had larger reduction in purchases (-9 percentage points) and consumption (-22 percentage points) (p=0.01)
Healthy Double Study	Polacsek et al. (2018)	RCT pilot	Determine if supermarket 2 for 1 on fruit and vegetables (FV) increases purchases among low-income families	USA, rural community in Portland, Maine 1 store	N = 354 Low income, 80% female, children under 18 years,	7 months (3 months baseline, 4 months post-intervention	Loyalty card transactions	N/A	Weekly sales ($) eligible FV	Linear regression	Intervention arm increased purchases of all FV (15%), fresh (18%), vegetables (20%), but no increase for fruit, little or negative effect for frozen and canned FV, vs control SNAP participants increased FV by 45%, vs 11% non-SNAP 53% increase in fresh,

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
											vs 13% non-SNAP
Targeted coupons	Guan et al. (2018)	Quasi-experimental	Influence of individually-targeted coupons on purchasing patterns for less healthful and more healthful products	USA 5 stores from same supermarket chain	N = 2,500 Convenience sample Demographic characteristics not described	2 years (2003 – 2005)	Loyalty card transactions collected by EPOS provider Dunnhumby	USDA Quarterly Food-At-Home Price Database (QFAHPB) used to categorise products	Weekly purchases (units) of 12 ‘healthful’ and ‘less healthful’ categories	Difference in difference analysis ANOVA	Weekly purchases increased from pre-post intervention periods for both exposed and unexposed. Exposed purchased 5.06 units more p<0.001 Positive difference in difference for all 12 groups, greatest for less healthful; convenience foods +1.17 units, lowest = nuts +0.03 units
Community interventions											
Shop Smart 4 Health	Ball, Kylie et al. (2016)	RCT	Cost-effectiveness of skill-building to promote purchase & consumptio	Australia, Coles stores Number of stores not stated	Low income women, regular shoppers at stores in deprived areas	12 months (6 months intervention + 6 months follow up)	Loyalty card transactions Self-reported portions/day, FFQ for	N/A	Purchases of veg & fruit (g/hh/wk), consumption (serves/day)	Generalised Estimating Equations, Mediation analysis & cost-consequence analysis	0.49 (CI 0.25, 0.72) portions more veg consumed immediately after, at 6-months +0.28 serves/day (CI 0.04, 0.52), ICER =

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
			n of fruit & veg				past 6 months			(broad societal perspective)	$3.10/extra veg serve/person/day
Supermarket Healthy Options Project (SHOP)	Eyles, H. et al. (2010)	Feasibility study	Develop culturally tailored nutrition education resources	New Zealand 6 stores	N = 551 Maori = 123, Pacific = 52, European /other = 346	3 months baseline purchase data used to inform design of materials	Store self-scan transaction data Australian Heart Foundation Tick nutrient profile to identify 'healthier' products	Supermarket Food & Nutrition Database (SFND); Manufactured Food Database, brand websites, back of pack & NZ national food tables	Feasibility of applying nutrient profiling	N/A	1814 (60%) products classified as 'healthier' Food & nutrient database successfully linked to 3 months transaction data Monthly reports automatically generated tailored shopping lists based on purchases
1% or Less campaign	Reger, B. et al. (1998)	Controlled community intervention	Effect of community education + mass media encouraging low fat milk consumption to reduce saturated fat intake	West Virginia, USA 2 intervention communities, 1 control (convenience sampled)	N = 25,000 in intervention communities N = 34,000 in control community	3 months (February – April 1995)	Monthly supermarket sales, & hh intake from telephone interviews	N/A	Mean sales milk (gallons)/supermarket/month (whole milk, 2% fat, 1%, ½% & skim) Market share by category (% total gallons) Self-reported consumption	Repeated measures ANOVA 2-tailed t-tests and F-tests at 5% significance level	Sales increased by 16%, low fat share up 23%. Greatest increase for 1% milk. Decrease in high fat milk, similar in control 38% respondents switched to low-fat, no difference by group, or individual characteristics.

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
	Reger, Bill et al. (1999)		Effect of mass media to change milk consumption	Wheeling, West Virginia USA	Population 35,000	6 weeks (February – March 1996)			, % high/low fat milk drinkers and switching		Low fat milk share increased17%, high fat decreased 13%. No difference in overall milk sales
	Reger, B. et al. (2000)		Advertising vs PR & community education	Rural West Virginia, USA	PR + education (n = 34,000), Advertising (n = 18,000), Control (n = 14,000)	8 weeks in winter 1997					PR + education: 19.6% switched to low-fat, 12.8% advertising, 6.8% control.
Towards a Healthy Diet	Dunt et al. (1999)	Quasi-experimental	Promote healthy diet policy changes in schools, health services, restaurants etc.	Victoria Australia 2 cities; 5 intervention stores, 4 control	N = 1137 completed panel questionnaires N = 703 completed cross-sectional survey	2 years (October 1991 – October 1993)	Monthly supermarket sales Panel and cross-sectional surveys	N/A	Sales volume of milk & table spreads /supermarket /month Self-reported opinion, dietary behaviour, cognition about healthy diet etc.	Mann-Whitney U-test for survey evaluation Method not stated for sales data	Modest positive changes in individuals - only significant group difference between = decrease in takeaway foods in intervention. No downward trend in unhealthy purchases.
In-store choice architecture											

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
Project Sol	Winkler et al. (2016)	Non-randomised intervention	Examine customer attitudes & sales effects of healthy checkout supermarket intervention	Denmark 4 chains owned by Coop group 28 stores (4 intervention, 12 control, 12 other areas)	Customers of different supermarket chains. Customer demographics unknown	5 months, 4 weeks intervention	Weekly store sales	N/A	Weekly store sales (revenue); all foods	Linear mixed models	Positive effect on carrot snack pack sales, but no other healthy snacks or fruit. Confectionary sales unaffected.
	Toft et al. (2017)	Non-randomised cluster intervention	Effect of improved shelf-space with & without 20% price discount for fruit & veg	Denmark 5 discount stores (2 intervention, 3 control) in 2 regions	Customer demographics unknown	5 months (1-month pre-, 3 months intervention, 1-month post)	Sales from Netto stores Sales from other intervention area supermarkets		Weekly store sales (units); fruit & veg (fresh, frozen, canned and dried); Index relative to previous year	Multi-level regression analysis	Shelf-space + price increased fresh by 22%, organic fresh by 12.1%, total by 15.3% No effect for shelf-space only No unhealthy substitution effects
Omega-3 podcasts	Bangia et al. (2017)	Non-randomised experiment	Impact of store podcast tour intervention on omega-3-rich foods	USA, New Jersey 20 stores in middle- & upper-middle class areas	N = 173 Loyalty customers who listened to n-3 podcasts during a main shop & shopped at least once/month	12 months (6 months pre-, 6 months post-intervention, 1 day)	Daily store loyalty card data	N/A	Sales of targeted n-3-rich foods (units/participant/ month by food type & category)	Pearson's correlation (intention & purchase) Wilcoxon signed-rank – pre-post differences by gender, SNAP participation & food type/ category	59% of shoppers increased n-3-rich purchases. Mean items significantly increased from 0.2 (SD 0.7) to 3.6 (SD 5.1) Increase in fortified foods greater for women than men (+2.68 items). No other

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
					during study period					Kruskal-Wallis 1-way ANOVA – pre-post differences by race or education	demographic differences No relationship between intention & purchases
Keyhole campaign	Mork et al. (2017)	Before and after	Impact of Keyhole awareness campaign on purchases of Keyhole-labelled products	Denmark 6 stores from 3 chains (2 regular, one discount chain)	Target = men >35-years with low education Data for all customers, no demographic information	9 weeks (3 weeks pre- and 3 weeks post-intervention)	Transaction data In store observation & researcher interviews	N/A	Daily store sales of 10 food categories – turnover (volume & value) by category & Keyhole status	Multi-level logistic regression	Odds of purchasing Keyhole labelled products rose by 20% in standard stores, 10% decrease in discount stores. Purchase more likely linked to health motives among participants with short education.
POP intervention	Freedman and Connors (2010)	Quasi-experimental pilot study	Effect of shelf tags to promote healthy choices	USA 1 on-campus convenience store at large urban university	Number unknown No customer demographics University students; 23% Asian, 16% Hispanic, 29% White, 32% Other	11 weeks (6 weeks fall 2008 semester + 5 weeks spring 2009)	Sales from computerised cash register	On pack nutritional information used to allocate tags indicating healthy food choice	Sales of tagged (healthy) & untagged (unhealthy) foods in 4 categories; cereal, soup, crackers, bread	Mann-Whitney U test at 95% confidence level	Increased sales of tagged items during intervention for cereal, soup & crackers but decrease for bread. Overall sales of tagged items increased 3.6% (SD 1.6%) P = 0.082

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
Shopper marketing intervention	Payne et al. (2015)	Non-randomised experiment	Efficacy of shopper marketing on produce demand, store profits & shopper budgets	El Paso, Texas, USA 4 stores (3 intervention, 1 control)	No customer demographics Area-level demographics;95% Hispanic, 53% female, Mean age ~30 years	Pilot = 14 days, main study = 28 days intervention + baseline & follow up 2012 - 2013	Store-level aggregated sales data	N/A	Total produce spend/person/ day, proportion of baseline & total expenditure (%)	T-test	Pilot: Significant increase in produce spend in intervention (+16%), not control (+4%). No change in overall spend. Main study: Both stores increased spend (+12.4% & +7.5%). proportion of total spend increased (+13.3% + 8.5%). No change in overall spend.
Manger Top intervention	Gamburzew et al. (2016)	Difference in differences	Social marketing to draw attention to inexpensive healthy foods	Marseilles, France 2 disadvantaged areas 4 discount stores (DIA)	Purchase data N = 6,625 loyalty card holders Survey subset (N = 116); 78% female, 16% food insecure, 31% aged >60years	18 months (January 2013 – June 2014), 6-month intervention (January-June 2014)	Loyalty card transactions In-depth survey	French food composition database	Contribution of inexpensive healthy foods to total food spend (%) & spend by category	Generalised linear model Chi-squared test Fisher tests One-way ANOVA	Contribution to total food spend ~20% for both groups, increased in 2014. No significant difference overall but greater increase in fruit, veg & starches for intervention stores
Eat Right 'N' Live Well!	Surkan et al. (2016)	Non-randomised	Multifaceted supermarket interventio	Baltimore, USA 2 stores in low-income	Customer number unknown.	3 years (July – October for 2010,	Store-level aggregated sales	N/A	Number of items (units) sold, absolute & %	Difference-in-difference analysis	Higher growth in sales of promoted foods in intervention store

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
		intervention	n promoting healthier alternatives to commonly purchased foods	African-American areas	Area demographics, 76% African-American, 20% unemployed, 33% single-parent hhs	2011 and 2012)			differences in sales		(+10.8%) vs control (+9.3%) Moderate success, not uniform across food categories
Guiding Stars	Hobin et al. (2017)	Natural experiment	Impact of Guiding Stars shelf labelling on nutritional quality of food purchases	Ontario, Canada 126 stores, 3 supermarket chains owned by the same company	Customer number unknown, 145 million transactions 783 exit interview participants; 75.0% female, 47.6% overweight/obese, 83.3% White	1 year (June 2012 – July 2013) Guiding Stars in 1 chain in August 2012	Aggregated supermarket transactions/day	Guiding Stars Licensing Co food and nutrient database – UPC-level, >55,000 products	Change in stars /product & /serve Change in calories & nutrients /serve Quantity of products /transaction, price/product, store revenue	Difference-in-difference analysis Regression analysis, controlling for seasonality	Significant increase in mean star rating (+1.4%), share of 1- and 3-star products (+2% & 1.9%), decline in 0-2-star products (-0.7% & -1.9%). 3.5% & 1.5% decrease in trans-fat & sugar, & 0.6% & 4.5% increase in fibre & omega-3 Number of products, price /product & store revenue increased. Effect varied by category.
Portion intervention	Vandenbroele et al. (2018)	Non-randomised	Effect of 2 additional smaller	Belgium 9 stores from large	N = 161 Loyalty card who	1 month	Supermarket aggregated	N/A	Unit sales & volume (kg) /store &	2-way ANOVA – before & during,	Slightly higher sales of two new smaller portions

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
		controlled experiment	portion options on portions purchased meat sausages	European retailer (1 intervention, 8 control)	customers bought target product 59% female. No demographics for baseline or control		& individual-level sales		/individual for target product (3 portions) & meat category	experiment vs control stores Control for backfire (purchasing multiple) & compensation within meat category	(52% combined) vs original (48%). Reduced total sales volume (kg). Small portion customers bought significantly less (kg) (M = 0.33, SD 1.90) than large customers (M = 0.49, SD = 1.91). No compensation
Make it Fresh for Less	Moran et al. (2019)	Quasi-experimental and RCT	Effect of healthful low-cost meal bundles (quasi-experimental) and electronic reminders (RCT) on purchases of healthy meal bundle items	USA 2 stores from large supermarket chain in Portland, Maine	N = 238 in RCT, intervention = 126, control = 112 81% female, 90% non-Hispanic White, 25% used SNAP benefits	13 months 40 weeks baseline, 16 weeks intervention	Store sales and loyalty card transactions	N/A	Sales ($) of meal bundle items, by transaction, or monthly by store	Linear regression	No effect of electronic reminders on purchases of meal bundle items. No significant increase in sales of bundled items in intervention store vs control.
Checkout nudges	Kroese et al. (2016)	Non-randomised controlled experiment	Investigate effect of a food repositioning nudge	Netherlands 3 kiosks (small convenienc	N = 91 participated in exit interviews	2 weeks (1 week baseline, 1 week nudge	Store-level daily sales (items)	N/A	Daily sales of nudged 'healthier' snacks	ANCOVA	Significant difference in mean daily number of nudged

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
			on healthy food choice	e stores) at train station	52% male, mean age 39 yrs (SD 15.75 yrs) Demographics of overall customer-base unknown	intervention)			(number of units)		items sold between stores Control = 23 items Nudge = 41 (p=0.00), Nudge + disclosure = 35 (p=0.02) No difference between nudge and nudge + disclosure (p=0.17) No difference in sales of non-nudged items
	Van Gestel et al. (2018)	Longitudinal		Netherlands 1 kiosk (small convenience store) at train station	N = 186 participated in exit interviews 57% male, mean age 38 yrs (SD 17 yrs) Demographics of overall customer-base unknown	8 weeks (4 weeks baseline, 4 weeks nudge intervention)	Store-level daily sales Individual-level purchases evaluated in exit interviews	N/A	Daily sales of selected healthy products as a proportion of total food sales	ANOVA	179 food products sold. Sales of total and healthy food products higher during baseline. Proportion of targeted healthy foods sold higher in nudge phase (mean = 6.3% SD 1.4) than baseline (mean = 4.3% SD 0.9) Effect maintained over 4-week nudge period

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
University store choice architecture	Walmsley et al. (2018)	Natural experiment	Effect of choice architecture intervention, re-arrangement of produce to increase the accessibility of fruit and vegetables	UK, Warwick 1 store on University Campus	Number unknown University students (26,000) and staff	5.5 years, excluding non-termtime weeks 90 weeks baseline, 40 weeks intervention A, 40 weeks intervention B	Store level sales data aggregated weekly	N/A	Fruit and vegetable (FV) sales (units and monetary spend) as a proportion of total food sales	Retrospective interrupted timeseries modelling Dynamic regression with Auto Regressive Integrated Moving Average (ARIMA)	Significant increase in proportion of FV for intervention A. Non-significant increase for intervention B. Overall downward trend in proportion of sales that were FV over the 5.5 year study period.
Online supermarket	Martinez et al. (2018)	Mixed methods RCT	Examine impact of pilot for online grocer to accept Electronic Benefit Transfer (EBT) Cards	USA Low-income neighbourhood in the Bronx	N = 148 included in baseline data 1/3 = EBT users N = 348 recruited to RCT	9 months September 2012 – June 2013	Online grocery transactions	N/A	Average spend per order (% of purchase) on 5 food groups; fruit, veg, dairy, sweets, salty snacks	Mann-Whitney U test	EBT orders spent more on sweets (10.8% vs 4.9% non-EBT) and salty snacks (2.2% of purchase, vs 1.1%), slightly less on fruit (6.3% vs 8.7%) . No significant difference for spend on veg (11.5% vs 14.2%) and dairy (both groups 8.5%)
Comparison with dietary intake											

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
Supermarket Healthy Options Project (SHOP)	Hamilton et al. (2007)	Observational	Compare supermarket nutrient availability with national consumption & expenditure surveys	New Zealand 1 store	N = 882 customers eligible for SHOP pilot RCT Age (µ) 38yrs, 73% female	12 months (February 2003 – January 2005)	Store self-scan (Shop 'N Go) transactions National consumption & expenditure surveys	Supermarket Food and Nutrition Database (SFND); Manufactured Food Database, brand websites, back of pack & NZ food tables	Proportion sales volume (units) & expenditure (%) by food category/hh macro-nutrients % energy Contribution of food groups to macro-nutrients	Difference between supermarket and survey data	Similar to survey for CHO, total fat & saturates, protein lower. Less comparable with children's survey (supermarket lower CHO, similar protein & saturates, & higher total fat). Expenditure similar for most foods, supermarkets lower for sweet foods & beverages

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
	Eyles, Helen et al. (2010)	Observational	Household electronic sales data vs individual nutrient intakes from 24-hr recalls	New Zealand 6 stores	N = 49 participants from SHOP RCT Age (µ) 48yrs, 84% female, 53% university/ tertiary qualifications	3 months (Nov 2004 – Jan 2005)	Self-scan (Shop 'N Go) transactions – coded to (SFND) 3000 top-selling foods 4 non-consecutive dietary recalls – coded to national food composition database (>2600 foods)		Household energy (E) & energy-adjusted macronutrients Energy density (ED) (beverages & non-beverages)	Spearman correlation coefficients Paired t-tests 2-sided at 5% significance level	Moderate correlation: Saturates (%E) R^2 = 0.54**, CHO (%E) R^2 = 0.48**, Protein (%E) R^2 = 0.44**, Fat (%E) R^2 = 0.34, Sugar (%E) R^2 = 0.33, ED nonbev (kcal/oz) R^2 = 0.37, ED bev (kcal/oz) R^2 = 0.09, Sodium (kcal/oz) R^2 = 0.06 No difference for saturates & total fat. Significant for; CHO +3%, Protein -4%, Sugar -2.1oz/kcal, Sodium -122.84, Oz/kcal
Population dietary surveillance											
USDA report	Frazao and Allshouse (1996)	Observational	Report the size & growth of USA nutritionally improved foods market	USA 3,000 supermarkets	No demographic information	5 years (1989 – 1993)	Store sales from supermarkets with annual revenue >$2 million	N/A	Volume sales, dollar sales, volume & dollar share of nutritionally improved products	Descriptive only, no statistical tests	Nutritionally improved cost more. Availability increased. Volume sales and dollar sales increased from 1989 to 1993

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
Columbus supermarket study	Jones (1997)	Observational	Difference in price elasticity for high & low-income customers	USA 7 stores	No customer demographic information. Supermarkets classified as high or low-income area based on census tract	2 x 54-weeks; 1990 - 1991 1993 - 1995	Weekly purchases by store	N/A	Price/ ounce (ratio of group sales) Elasticity of demand	Time-series cross-section regression model, error components model	Elasticity high for breakfast cereal, low for carbohydrates. Differences by income group for cereals. Lower income pay less per ounce & twice as price sensitive
Bread purchases	Revoredo-Giha et al. (2009)	Observational	Effect of price changes on consumption of different bread types in Scotland	Scotland; 3 TV regions (Borders, Central, North) Major UK supermarket	Number & demographics not reported. 3 geodemographic groups - proxy for affluence	2 years (Oct 2006 – Sept 2008)	Scanner data from loyalty customers	N/A	Regional & SES weekly sales premium & non-premium brown & white bread (g/person /day), (£/g)	3 demand models: Rotterdam demand system, Static LA/AIDS, Dynamic LA/AIDS	Brown & white bread quite price elastic - consumption reduces when prices increase, particularly brown bread. No difference in price elasticity by region or socioeconomic group
Finnish supermarket study	Närhinen et al. (1998)	Cross-sectional	Variation in daily sales, usefulness of supermarket data for monitoring & evaluating shopping behaviour	Finland 1 store in Mikkeli, town with ~30,000 inhabitants	All customers, no demographic information	2 months; May 1996, September 1996	Daily cash register sales aggregated weekly & monthly	N/A	Direct & proportional sales, 79 healthier & reference products, 17 categories (number, kg, price/kg)	Mean, standard deviation, coefficient of variation	Proportional more stable than direct sales. Variation similar for 1 week/1 month sales, greater daily variation – sales of milk & yoghurt higher on Fridays

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
	Närhinen et al. (1999)		How well does supermarket sales data reflect regional diet differences	Finland 8 Prisma stores, 6 cities		1 month, September 1997	Cash registers & 3 yrs health survey results (1995 – 1997)		Proportional sales milk, sour milk, fats, oils Mean salt & fat % & proportion saturated:total fat	Chi-squared	Regional differences in sales & survey data, high similarity (value not stated). Reported use of non-fat milk higher than actual sales.
Dalby CVD campaign	Radimer and Harvey (1998)	Cross-sectional	Validity of self-reported use of reduced fat & salt foods	Australia Remote community Dalby, 1 supermarket store	453 questionnaire respondents, no demographic information	1 year, 1992 - 1993	Sales data & FFQ 1 yr Milk deliveries data; 2 months 1993	N/A	% reporting use of reduced fat & salt foods Sales reduced fat & salt foods (adjusted for national adult milk consumption) Milk deliveries	Store sales within 91% of survey data (Y/N)	Reported consumption reduced fat & salt foods greater than sales & deliveries suggest Largest difference for reduced salt bread & soup, smallest for butter & margarine.
Study of Danish wine and beer drinkers	Johansen et al. (2006)	Observational	Investigate diet patterns of wine & beer buyers	Denmark 98 outlets; 2 large chains owned by Dansk Supermarked; 16	Customers of Bilka & Fotex, likely to over-represent middle income.	6 months (September 2002 – February 2003)	3.5 million transactions	N/A	Daily purchases of 40 food categories/ customer	Correspondence analysis & logistic regression	Wine buyers spent more & bought more items. More likely to follow Mediterranean diet – olives, fruit & veg, poultry, oil,

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
				Bilka, 82 Fotex	No customer demographic information.						& low-fat cheese, milk, meat Beer buyers follow traditional diet – ready meals, sugar, cold cuts, chips, pork, butter/margarine, sausages & lamb
Philadelphia supermarket	Phipps, Etienne J. et al. (2013)	Observational	Investigate predictors of fresh fruit & veg purchases in low income population	Philadelphia, USA 1 store, low-income minority ethnic community	30 low income loyalty card households, at least 1 child Primary household shopper; 90% female, 87% African-American, Mean age 42yrs (±14)	3 months, April 1 – June 30 2010	Loyalty card point of sale data	N/A	Primary = servings fresh fruit & veg/hh/week Secondary = total fresh produce expenditure/ hh/week	Bivariate & multivariable Poisson regression with log link	Controlling for household size, average servings +50-60% for each extra child (P=0.008), +10% for every year in age range of children (P=0.04) Mean servings/week = 4.0 (±2.9) No association with poverty, income, benefits, age or education of primary shopper
	Phipps, E. J. et al. (2014)		Impact of price discount on purchases of high-		82 primary household shoppers with loyalty card.	65 weeks; October 2012 – November 2013		Not stated	Weekly household sales of HCF & LCF, ratio odds of purchase,	Fixed effects logistic regression for ratio of odds Bivariate & multivariate	Odds of buying on sale vs full price higher for grain-based snacks, sweet snacks & SSBs (OR = 6.6,

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
			calorie foods (HCF) & low-calorie foods (LCF)		Adults ≥1 child, primarily female African-American				sale vs full price % cost saving/day	fixed effects generalised linear regression	5.9, 2.6 respectfully) all P<0.001. Not for savoury snacks or LCFs.
Swiss loyalty card study	Hauser et al. (2013)	Observational	Investigate how food-related values & attitudes influence purchases by category	Switzerland, 2 regions (German & French-speaking), 1 supermarket chain	851 loyalty card holders	Purchases 1 year prior to survey	Purchase data from loyalty card holders Values & attitudes survey	N/A	Annual hh expenditure/ category (% total food expenditure)	Confirmatory factor analysis Structural equation modelling	Moderate correlation between values & purchases. Fruit & veg associated with sustainability & health values. Fresh convenience & ready-to-eat positively correlated with convenience, negatively with conviviality & health.
Casino study	Hansel et al. (2015)	Observational	Relationship between purchases of alcoholic beverages & food	France, urban & rural areas Casino supermarkets	196,000 loyalty card holders, regular shoppers No demographic information	1 year (September 2010 – September 2011)	Purchase data from loyalty card holders	N/A	% of hh annual budget for alcohol, healthy & unhealthy food, ratio of total	k-means clustering by alcohol purchase chi-squared	Wine purchasers spent higher proportion of their budget on healthy foods vs beer or non-alcohol buyers. Non-alcohol buyers - lower total spend.

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
Healthy Food Program	Sturm et al. (2016)	Observational	Relationship between seasonal food purchases & BMI	South Africa	25% rebate (N = 400,000 households) BMI data for ~ 500,000 individuals	4 years (2009 – 2013)	Purchases from eligible supermarkets using health insurance credit card Health Risk Assessment Survey (HRA)	N/A	Monthly spend/category/hh : total spend monthly/hh	Multiple regression analysis at household (purchases) & individual level (BMI)	13% expenditure fruit & veg, 9% other healthy foods. December; 41% higher purchases of less desirable foods, lower fruit & veg purchases (vs January) Annual weight gain +0.13 BMI units (men +0.43kg, women +0.3kg). Christmas weight gain ~60-70% of annual (men +0.1 BMI units, +0.35kg, women +0.8 BMI units, +0.2kg)
Healthy Trolley Index	Taylor et al. (2015)	Observational	Healthy Trolley Index (HETI) to estimate diet quality & compare purchases with dietary guidelines	Australia Staff from large retailer corporate office wellness scheme,	964 loyalty card holders; mean age 37.6 yrs (SD 9.3), 56% female, mean BMI 27.9 (SD 6.6), 23% overweight, 28% obese,	1 month (April – May 2014)	Purchase data from loyalty card holders	N/A	HETI (/100) high = compliance Australian Guide to Healthy Eating (AGHE) Expenditure/ HETI group, proportion	One-way ANOVA & Bonferroni post-hoc tests Chi-squared difference in shopping frequency by weight	Average HETI = 58.8 (SD 10.9), higher for males & normal BMI 62.7% met meat & alternatives guidance, compliance poor for grains (0.2%), discretionary (1.8%) & veg (5%).

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
					15% live alone				total food & drink		
SNAP monthly purchases	Francklet et al. (2019)	Secondary analysis of RCTs	Examine purchase fluctuations over the SNAP benefit month, for SNAP households and non-SNAP households	USA 2 same chain supermarkets in low-income communities in Maine	N=950 loyalty card holders who participated in RCTs 84% female, 94% White non-Hispanic	Up to 8 months of data	Daily purchases aggregated by week and by month, from loyalty card issued for RCT	N/A	Mean spend ($) per transaction on all foods and for selected categories in first 2 weeks, and last 2 weeks of the month after SNAP benefits issued	Difference-in-difference	37% decline for SNAP (all categories), 3% for non-SNAP (only red meat and poultry) SNAP decline by category; veg -25%, fruit -27%, SSBs -30%, red meat - 37%, convenience -40%, poultry -48%
LoCard	Nevalainen et al. (2018)	Longitudinal	Address potential and challenges of loyalty card data for health research.	Finland, 1 grocery chain	N = 14,595 households >13 million transactions	1 year (1 January – 31 December 2016)	Loyalty card data	N/A	Total annual grocery expenditure (Euros) Participant and purchasing profiles	Descriptive statistics, linear regression, logistic regression, inverse probability weighting to adjust for non-participation bias	Gender and age are significant determinants of expenditure (peak at middle age). Men's expenditure greater than women's. Top 10 selling product groups included 'beer and cigarettes'. Fridays and Saturdays = most active purchase days. National

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
											holidays preceded by peak in spend
	Uusitalo et al. (2019)	Longitudinal	analyse alcohol purchase patterns on the level of individual shopping occasions.		N = 13,274 households				Total expenditure (Euros) & proportion of total basket expenditure on alcohol (beer, cider, non-alcoholic equivalents), cigarettes & food groups	K-means cluster analysis based on alcohol purchases Linear mixed models to assess difference in means between clusters	8 clusters, most common = no alcohol (86.1%) (reference) Beer buyers mostly men, and older. More alcohol associated with more food, especially meat, soft drinks, cheese, sweet foods, fat, breads, ready to eat. But lower fruit expenditure.
Methodological											
Informatics feasibility study	Brinkerhoff et al. (2011)	Feasibility study	Feasibility of linking point-of-sale data to USDA-SR nutrient database	USA, Intermountain West Large supermarket chain	32,785 customers 2,009,533 de-identified sales items No demographic information	2 weeks, August 2007	Individual customer purchase data	United States Department for Agriculture National Nutrient Database for Standard Reference (USDA-SR)	Match-rate (%) between product in supermarket sales database and USDA-SR	No statistical analysis methods String-matching & fuzzy matching at the food item level, manual matching	3-tier organisational hierarchy (Department → Commodity → Sub-commodity) 70% sub-commodities mapped to SR food items (complete nutritional data), 100% of sub-

Study	Author Year	Design	Aim	Setting	Population	Duration	Data source	Nutrition data source	Outcomes	Analysis	Key findings
								>7,500 food items, 24 food groups			commodities mapped to SR food groups
qDIET	Chidambaram et al. (2013)	Feasibility study	qDIET method - automated & self-sustaining linkage between retail data & USDA databases to calculate HEI	USA, Salt Lake City Large national grocery retailer	50 households who reported >75% food intake from retail stores No demographic information	12 months (February 2007 – April 2008)	Household purchase data for loyalty card holders	Food & Nutrition Database for Dietary Studies (FNDDS), MyPyramid Equivalents Database (MPED), Google Shopping API & Factual.com API	Match-rate (%) between product in supermarket sales database & API product data	Non-parametric Kolmogorov-Smirnov empirical distribution function two-sample test	No significant difference in HEI scores distribution between retail households & NHANES 30.7% of the 12,332 products matched by Google API, 71.5% matched by Factual
FPED	Tran et al. (2017)	Feasibility study	Systematic food quality monitoring by automated mapping to USDA Food Patterns Equivalent Database (FPDB)	USA, 4 geographic regions 1 grocery chain	144,000 households 190 million transactions 92,062 distinct grocery items	15 months (January 2012 – March 2013)	Household purchase data	USDA databases; Food & Nutrient Database for Dietary Studies (FNDDS), Food Pattern Equivalents Database (FPED)	Match-rate (%) between product in supermarket sales database & FPED	Confidence coefficient for similarity between grocery description & food FPED categories	Match-rate per category between 77% - 100% Mappings more complex for mixed dishes & ethnic foods – yet to be verified

Table S5 – NIH Quality Assessment Tool for Observational Cohort and Cross-sectional Studies Risk of Bias assessment for included papers

Author, year	1. Question clear?	2. Population clearly defined?	3. >50% participate?	4. Recruitment populations consistent?	5. Sample size justified?	6. Exposure assessed prior to outcome?	7. Sufficient timeframe?	8. Different exposure levels?	9. Repeated exposure assessment?	10. Valid outcomes?	11. Outcome assessors blinded?	12. Loss to follow up <20%?	13. Confounders adjusted for?	Overall quality rating
Alvarado et al. (2019)	✓	✓	?	?	X	✓	✓	X	✓	✓	NA	?	✓	Fair
An (2014)	✓	✓	?	✓	X	✓	✓	✓	✓	✓	NA	?	X	Fair
Andreyeva and Luedicke (2013)	✓	✓	?	✓	X	✓	✓	X	✓	✓	NA	?	✓	Fair
Andreyeva and Luedicke (2015)	✓	✓	?	✓	X	✓	✓	✓	✓	✓	NA	?	✓	Fair
Andreyeva et al. (2014)	✓	✓	?	✓	X	✓	✓	✓	✓	✓	NA	?	✓	Fair
Andreyeva et al. (2012)	✓	✓	?	✓	X	✓	✓	✓	✓	✓	NA	?	✓	Fair
Andreyeva et al. (2013)	✓	✓	?	✓	X	✓	✓	X	✓	✓	NA	?	✓	Fair
Balasubramanian and Cole (2002)	✓	?	?	?	X	✓	✓	✓	✓	✓	NA	?	X	Poor
Ball, Kylie et al. (2016)	✓	✓	X	✓	X	✓	✓	X	✓	✓	✓	✓	X	Fair
Ball, K. et al. (2015)	✓	✓	✓	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	Good
Banerjee and Nayak (2018)	✓	✓	?	✓	✓	✓	X	✓	✓	X	?	✓	✓	Poor
Bangia et al. (2017)	✓	✓	✓	✓	X	✓	✓	✓	X	✓	NA	X	✓	Fair

Bartlett (2014)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	?	X	✓	Good
Bernales-Korins et al. (2017)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	?	X	✓	Good
Blakely et al. (2011)	✓	✓	✓	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	Good
Brimblecombe et al. (2017)	✓	X	?	✓	✓	✓	✓	✓	✓	✓	?	?	X	Fair
Brinkerhoff et al. (2011)	✓	X	?	?	X	✓	NA	NA	NA	✓	NA	?	NA	Poor
Brunello et al. (2012)	✓	X	?	✓	X	✓	✓	✓	✓	✓	X	?	✓	Poor
Brunello et al. (2014)	✓	X	?	✓	X	✓	✓	✓	✓	✓	X	?	✓	Poor
Chidambaram et al. (2013)	X	X	?	?	X	✓	✓	✓	✓	?	NA	✓	NA	Poor
Dunt et al. (1999)	✓	X	?	?	X	✓	✓	✓	✓	?	NA	?	?	Poor
Eyles, H. et al. (2010)	✓	✓	✓	✓	✓	✓	✓	X	NA	✓	NA	✓	X	Fair
Eyles, Helen et al. (2010)	✓	✓	?	✓	X	NA	NA	NA	NA	?	NA	?	X	Poor
Ferguson et al. (2017)	✓	✓	?	?	X	✓	✓	✓	✓	✓	NA	?	✓	Fair
Franckle et al. (2018)	✓	✓	✓	✓	X	✓	✓	✓	✓	✓	✓	X	✓	Fair
Franckle et al. (2019)	✓	✓	?	✓	X	✓	✓	X	✓	✓	NA	?	✓	Fair
Frazao and Allshouse (1996)	✓	X	?	?	X	✓	✓	✓	✓	X	NA	?	X	Poor
Freedman and Connors (2010)	✓	✓	?	?	X	✓	?	✓	✓	✓	NA	?	✓	Fair
Gamburzew et al. (2016)	✓	✓	✓	✓	X	✓	✓	✓	✓	✓	NA	X	✓	Good
Geliebter et al. (2013)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	?	X	✓	Good
Guan et al. (2018)	✓	X	?	?	X	✓	✓	✓	✓	✓	NA	?	X	Poor
Hamilton et al. (2007)	✓	✓	✓	✓	X	NA	✓	NA	NA	✓	NA	✓	X	Fair
Hansel et al. (2015)	✓	✓	?	?	X	✓	✓	✓	✓	✓	NA	?	✓	Fair
Hauser et al. (2013)	✓	✓	?	?	X	✓	✓	✓	NA	✓	NA	?	✓	Poor

Hobin et al. (2017)	✓	✓	?	?	X	✓	✓	X	✓	✓	NA	?	X	Fair
Johansen et al. (2006)	✓	X	?	?	X	✓	✓	✓	✓	✓	NA	?	X	Fair
Jones (1997)	✓	?	?	?	X	✓	✓	✓	✓	✓	NA	?	X	Poor
Kroese et al. (2016)	✓	X	?	?	X	✓	✓	✓	X	✓	NA	?	X	Poor
Le et al. (2016)	✓	✓	✓	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	Good
Martinez et al. (2018)	✓	X	X	✓	X	✓	✓	✓	✓	✓	NA	X	?	Fair
Mathios (1998)	✓	✓	?	?	X	✓	✓	✓	✓	✓	NA	?	✓	Fair
Mathios (2000)	✓	✓	?	?	X	✓	✓	✓	✓	✓	NA	?	✓	Fair
Moran et al. (2019)	✓	✓	✓	✓	X	✓	✓	✓	✓	X	X	✓	X	Poor
Mork et al. (2017)	✓	X	?	?	X	✓	✓	✓	✓	✓	NA	?	?	Poor
Närhinen et al. (1998)	✓	X	?	?	X	✓	✓	✓	✓	✓	NA	?	?	Fair
Närhinen et al. (1999)	✓	?	?	?	X	✓	✓	✓	✓	✓	NA	?	?	Fair
Nevalainen et al. (2018)	X	✓	X	✓	X	✓	✓	✓	✓	✓	NA	✓	✓	Fair
Ni Mhurchu et al. (2007)	✓	✓	✓	X	X	✓	✓	NA	NA	✓	?	✓	?	Fair
Ni Mhurchu et al. (2010)	✓	✓	✓	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	Good
Payne et al. (2015)	✓	✓	?	?	X	✓	✓	✓	✓	✓	NA	?	?	Fair
Phipps, E. J. et al. (2014)	✓	✓	?	✓	X	✓	✓	✓	✓	✓	NA	?	✓	Fair
Phipps, Etienne J. et al. (2013)	✓	✓	?	✓	X	✓	✓	✓	✓	✓	NA	?	✓	Fair
Polacsek et al. (2018)	✓	✓	✓	✓	X	✓	✓	✓	✓	✓	✓	✓	✓	Good
Radimer and Harvey (1998)	✓	X	X	✓	X	✓	✓	✓	✓	✓	NA	?	X	Poor
Reger, B. et al. (2000)	✓	✓	?	✓	X	✓	✓	✓	✓	✓	NA	X	?	Fair
Reger, B. et al. (1998)	✓	X	?	?	?	✓	✓	X	✓	✓	NA	✓	✓	Fair
Reger, Bill et al. (1999)	✓	✓	?	?	X	✓	?	X	✓	✓	NA	?	X	Poor

Revoredo-Giha et al. (2009)	✓	X	?	?	X	✓	✓	✓	NA	✓	NA	?	?	Fair
Schwartz et al. (2014)	✓	✓	✓	✓	X	✓	✓	✓	✓	✓	?	✓	✓	Fair
Silver et al. (2017)	✓	✓	X	✓	✓	✓	✓	✓	✓	✓	X	X	✓	Fair
Stead et al. (2017)	✓	✓	?	✓	X	✓	✓	✓	✓	✓	?	✓	?	Fair
Sturm et al. (2013)	✓	✓	✓	✓	X	✓	✓	✓	✓	✓	X	?	✓	Fair
Sturm et al. (2016)	✓	✓	✓	✓	X	✓	✓	NA	✓	✓	NA	✓	✓	Good
Surkan et al. (2016)	✓	X	?	✓	X	✓	✓	✓	✓	✓	NA	?	?	Fair
Taylor et al. (2015)	✓	✓	?	✓	X	✓	✓	✓	✓	✓	NA	✓	✓	Fair
Toft et al. (2017)	✓	✓	?	?	X	✓	✓	✓	✓	✓	X	?	?	Fair
Tran et al. (2017)	X	X	?	?	?	✓	✓	✓	✓	✓	NA	?	NA	Poor
Uusitalo et al. (2019)	✓	✓	✓	?	X	✓	✓	✓	✓	✓	NA	?	✓	Fair
Van Gestel et al. (2018)	✓	X	?	?	X	✓	✓	✓	✓	✓	NA	?	X	Poor
Vandenbroele et al. (2018)	✓	X	?	?	X	✓	✓	✓	✓	✓	NA	?	✓	Fair
Walmsley et al. (2018)	✓	X	?	?	X	✓	✓	✓	✓	✓	NA	?	X	Poor
Winkler et al. (2016)	✓	✓	?	?	X	✓	?	✓	✓	✓	NA	?	?	Fair

✓ = Yes, X = No, NA = Not Applicable, ? = Not Reported or Cannot Determine

References

Aggarwal, A., Cook, A.J., Jiao, J., Seguin, R.A., Moudon, A.V., Hurvitz, P.M. and Drewnowski, A. 2014. Access to Supermarkets and Fruit and Vegetable Consumption. *American Journal of Public Health.* **104**(5), pp.917-923.

Alvarado, M., Unwin, N., Sharp, S.J., Hambleton, I., Murphy, M.M., Samuels, T.A., Suhrcke, M. and Adams, J. 2019. Assessing the impact of the Barbados sugar-sweetened beverage tax on beverage sales: an observational study. *International Journal of Behavioral Nutrition & Physical Activity.* **16**(1), p13.

Andreyeva, T. and Luedicke, J. 2013. Federal food package revisions: effects on purchases of whole-grain products. *American Journal of Preventive Medicine.* **45**(4), pp.422-429.

Andreyeva, T. and Luedicke, J. 2015. Incentivizing fruit and vegetable purchases among participants in the special supplemental nutrition program for women, infants, and children. *Public Health Nutrition.* **18**(1), pp.33-41.

Andreyeva, T., Luedicke, J., Henderson, K.E. and Schwartz, M.B. 2014. The positive effects of the revised milk and cheese allowances in the special supplemental nutrition program for women, infants, and children. *Journal of the Academy of Nutrition and Dietetics.* **114**(4), pp.622-630.

Andreyeva, T., Luedicke, J., Henderson, K.E. and Tripp, A.S. 2012. Grocery store beverage choices by participants in federal food assistance and nutrition programs. *American Journal of Preventive Medicine.* **43**(4), pp.411-418.

Arribas-Bel, D. 2014. Accidental, Open and Everywhere: Emerging Data Sources for the Understanding of Cities. *Applied Geography.* **49**, pp.45–53.

Balasubramanian, S.K. and Cole, C. 2002. Consumers' search and use of nutrition information: The challenge and promise of the Nutrition Labeling and Education Act. *Journal of Marketing.* **66**(3), pp.112-127.

Ball, K., McNaughton, S.A., Le, H.N., Abbott, G., Stephens, L.D. and Crawford, D.A. 2016. ShopSmart 4 Health: results of a randomized controlled trial of a behavioral intervention promoting fruit and vegetable consumption among socioeconomically disadvantaged women. *The American journal of clinical nutrition.* **104**(2), pp.436-445.

Ball, K., McNaughton, S.A., Le, H.N.D., Gold, L., Ni Mhurchu, C., Abbott, G., Pollard, C. and Crawford, D. 2015. Influence of price discounts and skill-building strategies on purchase and consumption of healthy food and beverages: Outcomes of the supermarket healthy eating for life randomized controlled trial. *American Journal of Clinical Nutrition.* **101**(5), pp.1055-1064.

Bandy, L., Rayner, M., Jebb, S. and Adhikari, V. 2018. *The use of food sales databases for public health nutrition research: a systematic review.* [Online]. Available from: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=91421

Banerjee, T. and Nayak, A. 2018. Believe it or not: Health education works. *Obesity Research and Clinical Practice.* **12**(1), pp.116-124.

Birkin, M., Wilkins, E. and Morris, M.A. 2019. Creating a long-term future for big data in obesity research. *International Journal of Obesity.* **43**(12), pp.2587-2592.

Blakely, T., Ni Mhurchu, C., Jiang, Y., Matoe, L., Funaki-Tahifote, M., Eyles, H.C., Foster, R.H., McKenzie, S. and Rodgers, A. 2011. Do effects of price discounts and nutrition education on food purchases vary by ethnicity, income and education? Results from a randomised, controlled trial. *Journal of Epidemiology and Community Health.* **65**(10), pp.902-908.

Bland, M.a.A., D. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet.* **327**(8476), pp.307 - 310.

Brimblecombe, J., Ferguson, M., Chatfield, M.D., Liberato, S.C., Gunther, A., Ball, K., Moodie, M., Miles, E., Magnus, A., Ni Mhurchu, C., Leach, A.J., Bailie, R. and collaborative, S.R.r. 2017. Effect of a price discount and consumer education strategy on food and beverage purchases in remote Indigenous Australia: a stepped-wedge randomised controlled trial. *The Lancet. Public health.* **2**(2), pp.e82-e95.

Brinkerhoff, K.M., Brewster, P.J., Clark, E.B., Jordan, K.C., Cummins, M.R. and Hurdle, J.F. 2011. Linking Supermarket Sales Data To Nutritional Information: An Informatics Feasibility Study. *AMIA Annual Symposium Proceedings.* **2011**, pp.598-606.

Brunello, G., Paola, M.d. and Labartino, G. 2012. More apples less chips? The effect of school fruit schemes on the consumption of junk food. *IZA Discussion Papers - Forschungsinstitut zur Zukunft der Arbeit.* (6496), pp.23-pp.

Brunello, G., Paola, M.d. and Labartino, G. 2014. More apples fewer chips? The effect of school fruit schemes on the consumption of junk food. *Health Policy.* **118**(1), pp.114-126.

Buttriss, J.L., Welch, A.A., Kearney, J.M. and Lanham-New, S.A. 2017. *Public Health Nutrition.* Wiley.

Carter, M.C., Hancock, Neil., Albar, Salwa A., Brown, Helen., Greenwood, Darren C., Hardie, Laura J., Frost, Gary S., Wark, Petra, A., and Cade, Janet E. 2016. Development of a New Branded UK Food Composition Database for an Online Dietary Assessment Tool. *Nutrients.* **8**(480).

Chidambaram, V., Brewster, P.J., Jordan, K.C. and Hurdle, J.F. 2013. qDIET: toward an automated, self-sustaining knowledge base to facilitate linking point-of-sale grocery items to nutritional content. *AMIA ... Annual Symposium proceedings. AMIA Symposium.* **2013**, pp.224-233.

Clark, S.D., Shute, B., Jenneson, V., Rains, T., Birkin, M. and Morris, M.A. 2021. Dietary Patterns Derived from UK Supermarket Transaction Data with Nutrient and Socioeconomic Profiles. *Nutrients.* **13**(5), p1481.

Cochrane Public Health Group. 2011. *Data Extraction and Assessment Template*. UK. Available from: https://ph.cochrane.org/sites/ph.cochrane.org/files/public/uploads/CPHG%20Data%20extraction%20template_0.docx

Data For Good Foundation. 2021. [Online]. [Accessed 13.05.2021]. Available from: https://dataforgoodfoundation.com/en/

Dunt, D., Day, N. and Pirkis, J. 1999. Evaluation of a community-based health promotion program supporting public policy initiatives for a healthy diet. *Health Promotion International.* **14**(4), pp.317-327.

Einav, L., Leibtag, Ephraim., Nevo, Aviv. 2008. *On the Accuracy of Nielsen Homescan Data.* U.S. Department of Agriculture.

Eyles, H., Jiang, Y. and Ni Mhurchu, C. 2010. Use of household supermarket sales data to estimate nutrient intakes: a comparison with repeat 24-hour dietary recalls. *Journal of the American Dietetic Association.* **110**(1), pp.106-110.

FAO. 2021. *International Network of Food Data Systems (INFOODS).* [Online]. [Accessed August 2021]. Available from: http://www.fao.org/infoods/infoods/en/

Ferguson, M., O'Dea, K., Holden, S., Miles, E. and Brimblecombe, J. 2017. Food and beverage price discounts to improve health in remote Aboriginal communities: mixed method evaluation of a natural experiment. *Australian and New Zealand journal of public health.* **41**(1), pp.32-37.

Franckle, R.L., Levy, D.E., Macias-Navarro, L., Rimm, E.B. and Thorndike, A.N. 2018. Traffic-light labels and financial incentives to reduce sugar-sweetened beverage purchases by low-income Latino families: a randomized controlled trial. *Public Health Nutrition.* **21**(8), pp.1426-1434.

Franckle, R.L., Thorndike, A.N., Moran, A.J., Hou, T., Blue, D., Greene, J.C., Bleich, S.N., Block, J.P., Polacsek, M. and Rimm, E.B. 2019. Supermarket Purchases Over the Supplemental Nutrition Assistance Program Benefit Month: A Comparison Between Participants and Nonparticipants. *American Journal of Preventive Medicine.* **57**(6), pp.800-807.

Fraser, L., Edwards, K., Cade, J. and Clarke, G. 2010. The Geography of Fast Food Outlets: A Review. *International journal of environmental research and public health.* **7**, pp.2290-2308.

Frazao, E. and Allshouse, J.E. 1996. Size and growth of the nutritionally improved foods market. *Agriculture Information Bulletin - United States Department of Agriculture.* (723), pp.iv-pp.

Freedman, M.R. and Connors, R. 2010. Point-of-Purchase Nutrition Information Influences Food-Purchasing Behaviors of College Students: A Pilot Study. *Journal of the American Dietetic Association.* **110**(8), pp.1222-1226.

Gamburzew, A., Darcel, N., Gazan, R., Dubois, C., Maillot, M., Tome, D., Raffin, S. and Darmon, N. 2016. In-store marketing of inexpensive foods with good nutritional quality in disadvantaged neighborhoods: Increased awareness, understanding, and purchasing. *The International Journal of Behavioral Nutrition and Physical Activity.* **13**.

Green, M.A., Watson, A.W., Brunstrom, J.M., Corfe, B.M., Johnstone, A.M., Williams, E.A. and Stevenson, E. 2020. Comparing supermarket loyalty card data with traditional diet survey data for understanding how protein is purchased and consumed in older adults for the UK, 2014–16. *Nutrition Journal.* **19**(1), p83.

Greenwood, D.C., Ransley, J.K., Gilthorpe, M.S. and Cade, J.E. 2006. Use of Itemized Till Receipts to Adjust for Correlated Dietary Measurement Error. *American Journal of Epidemiology.* **164**(10), pp.1012-1018.

Guan, X., Atlas, S.A. and Vadiveloo, M. 2018. Targeted retail coupons influence category-level food purchases over 2-years. *The International Journal of Behavioral Nutrition and Physical Activity Vol 15 2018, ArtID 111.* **15**(1), p111.

Hamilton, S., Ni Mhurchu, C. and Priest, P. 2007. Food and nutrient availability in New Zealand: An analysis of supermarket sales data. *Public Health Nutrition.* **10**(12), pp.1448-1455.

Hansel, B., Roussel, R., Diguet, V., Deplaude, A., Chapman, M.J. and Bruckert, E. 2015. Relationships between consumption of alcoholic beverages and healthy foods: The French supermarket cohort of 196,000 subjects. *European Journal of Preventive Cardiology.* **22**(2), pp.215-222.

Hauser, M., Nussbeck, F.W. and Jonas, K. 2013. The impact of food-related values on food purchase behavior and the mediating role of attitudes: A Swiss study. *Psychology & Marketing.* **30**(9), pp.765-778.

Hobin, E., Bollinger, B., Sacco, J., Liebman, E., Vanderlee, L., Zuo, F., Rosella, L., L'Abbe, M., Manson, H. and Hammond, D. 2017. Consumers' response to an on-shelf nutrition labelling system in supermarkets: Evidence to inform policy and practice. *Milbank Quarterly.* **95**(3), pp.494-534.

Jenneson, V., Clarke, G.P., Greenwood, D.C., Shute, B., Rains, T. and Morris, M.A. 2021. *Exploring the geographic variation in food purchasing behaviour using supermarket transaction data*. Unpublished.

Jenneson, V., Morris, M. A., Greenwood, D., Clarke, G., Pontin, F. 2018. *The use of electronic supermarket sales data for population dietary surveillance*. PROSPERO: CRD. [Systematic review registration]. Available from: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=103470

Jenneson, V., Shute, B., Greenwood, D., Clarke, G., Clark, S., Rains, T. and Morris, M.A. 2020. Variation in fruit and vegetable purchasing patterns in Leeds: using novel loyalty card transaction data. *Proceedings of the Nutrition Society.* **79**(OCE2), pE670.

Johansen, D., Friis, K., Skovenborg, E. and Grønbæk, M. 2006. Food buying habits of people who buy wine or beer: cross sectional study. *BMJ.* **332**(7540), pp.519-522.

Jones, E. 1997. An Analysis of Consumer Food Shopping Behavior Using Supermarket Scanner Data: Differences by Income and Location. *American Journal of Agricultural Economics.* **79**(5), pp.1437-1443.

Kitchin, R. and McArdle, G. 2016. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society.* **3**(1), p2053951716631130.

Kroese, F.M., Marchiori, D.R. and de Ridder, D.T. 2016. Nudging healthy food choices: a field experiment at the train station. *Journal of public health (Oxford, England).* **38**(2), pp.e133-e137.

Laney, D. 2013. *3D data management: Controlling data volume, velocity and variety.* [Online]. [Accessed 28.02.18]. Available from: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Le, H.N.D., Gold, L., Abbott, G., Crawford, D., McNaughton, S.A., Ni Mhurchu, C., Pollard, C. and Ball, K. 2016. Economic evaluation of price discounts and skill-building strategies on purchase and consumption of healthy food and beverages: The SHELf randomized controlled trial. *Social Science & Medicine.* **159**, pp.83-91.

Martinez, O., Tagliaferro, B., Rodriguez, N., Athens, J., Abrams, C. and Elbel, B. 2018. EBT payment for online grocery orders: A mixed-methods study to understand its uptake among SNAP recipients and the barriers to and motivators for its use. *Journal of Nutrition Education and Behavior.* **50**(4), pp.396-402.

Mathios, A.D. 1998. The Importance of Nutrition Labeling and Health Claim Regulation on Product Choice: An Analysis of the Cooking Oils Market. *Agricultural and Resource Economics Review.* **27**(2), pp.159-168.

Mathios, A.D. 2000. The Impact of Mandatory Disclosure Laws on Product Choices: An Analysis of the Salad Dressing Market. *The Journal of Law & Economics.* **43**(2), pp.651-678.

Mooney, S.J. and Pejaver, V. 2018. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annual Review of Public Health.* **39**(1), pp.95-112.

Moran, A.J., Khandpur, N., Polacsek, M., Thorndike, A.N., Franckle, R.L., Boulos, R., Sampson, S., Greene, J.C., Blue, D.G. and Rimm, E.B. 2019. Make It Fresh, for Less! A supermarket meal bundling and electronic reminder intervention to promote healthy purchases among families with children. *Journal of Nutrition Education and Behavior.* **51**(4), pp.400-408.

Mork, T., Grunert, K.G., Fenger, M., Juhl, H.J. and Tsalis, G. 2017. An analysis of the effects of a campaign supporting use of a health symbol on food sales and shopping behaviour of consumers. *BMC public health.* **17**(1), p239.

Morris, M., Clarke, G., Edwards, K., Hulme, C. and Cade, J. 2016. Geography of Diet in the UK Women's Cohort Study: A Cross-Sectional Analysis. *Epidemiology - Open Journal.* **1**(1), pp.20 - 32.

Morris, M., Wilkins, E.L., Galazoula, M., Clark, S.D. and Birkin, M. 2020. Assessing diet in a university student population: A longitudinal food card transaction data approach. *British Journal of Nutrition.* **123**.

Morris, M.A., Wilkins, E., Timmins, K.A., Bryant, M., Birkin, M. and Griffiths, C. 2018. Can big data solve a big problem? Reporting the obesity data landscape in line with the Foresight obesity system map. *International Journal of Obesity.* **42**(12), pp.1963-1976.

MRC. 2017. *Review of Nutrition and Human Health Research.* [Online]. Available from: https://mrc.ukri.org/documents/pdf/review-of-nutrition-and-human-health/

Närhinen, M., Berg, M.-A., Nissinen, A. and Puska, P. 1999. Supermarket sales data: a tool for measuring regional differences in dietary habits. *Public Health Nutrition.* **2**(3), pp.277-282.

Närhinen, M., Nissinen, A. and Puska, P. 1998. Sales data of a supermarket – a tool for monitoring nutrition interventions. *Public Health Nutrition.* **1**(2), pp.101-107.

Nevalainen, J., Erkkola, M., Saarijärvi, H., Näppilä, T. and Fogelholm, M. 2018. Large-scale loyalty card data in health research. *Digital health.* **4**, pp.2055207618816898-2055207618816898.

Ni Mhurchu, C., Blakely, T., Jiang, Y., Eyles, H.C. and Rodgers, A. 2010. Effects of price discounts and tailored nutrition education on supermarket purchases: A randomized controlled trial. *American Journal of Clinical Nutrition.* **91**(3), pp.736-747.

Ni Mhurchu, C., Blakely, T., Wall, J., Rodgers, A., Jiang, Y. and Wilton, J. 2007. Strategies to promote healthier food purchases: A pilot supermarket intervention study. *Public Health Nutrition.* **10**(6), pp.608-615.

NIH. 2017. *Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies*. National Heart, L., and Blood Institute. NHLBI. Available from: https://www.nhlbi.nih.gov/health-pro/guidelines/in-develop/cardiovascular-risk-reduction/tools/cohort

Payne, C.R., Niculescu, M., Just, D.R. and Kelly, M.P. 2015. Shopper marketing nutrition interventions: Social norms on grocery carts increase produce spending without increasing shopper budgets. *Preventive Medicine Reports.* **2**, pp.287-291.

Phipps, E.J., Kumanyika, S.K., Stites, S.D., Singletary, S.B., Cooblall, C. and DiSantis, K.I. 2014. Buying food on sale: a mixed methods study with shoppers at an urban supermarket, Philadelphia, Pennsylvania, 2010-2012. *Preventing Chronic Disease.* **11**, pE151.

Phipps, E.J., Stites, S.D., Wallace, S.L. and Braitman, L.E. 2013. Fresh fruit and vegetable purchases in an urban supermarket by low-income households. *Journal of Nutrition Education and Behavior.* **45**(2), pp.165-170.

Polacsek, M., Moran, A., Thorndike, A.N., Boulos, R., Franckle, R.L., Greene, J.C., Blue, D.J., Block, J.P. and Rimm, E.B. 2018. A Supermarket Double-Dollar Incentive Program Increases Purchases of Fresh Fruits and Vegetables Among Low-Income Families With Children: The Healthy Double Study. *Journal of nutrition education and behavior.* **50**(3), pp.217-228.

Popay, J., Roberts, Helen., Sowden, Amanda., Petticrew, Mark., Arai, Lisa., Rodgers, Mark., Britten, Nicky., Roen, Katrina and Duffy, Steven. 2006. *Guidance on the Conduct of Narrative Synthesis in Systematic Reviews*. Programme, E.M. UK: University of Lancaster. [Guidance document].

Radimer, K.L. and Harvey, P.W. 1998. Comparison of self-report of reduced fat and salt foods with sales and supply data. *European journal of clinical nutrition.* **52**(5), pp.380-382.

Ransley, J.K., Donnelly, J.K., Botham, H., Khara, T.N., Greenwood, D.C. and Cade, J.E. 2003. Use of supermarket receipts to estimate energy and fat content of food purchased by lean and overweight families. *Appetite.* **41**(2), pp.141-148.

Ransley, J.K., Donnelly, J.K., Khara, T.N., Botham, H., Arnot, H., Greenwood, D.C. and Cade, J.E. 2001. The use of supermarket till receipts to determine the fat and energy intake in a UK population. *Public Health Nutrition.* **4**(6), pp.1279-1286.

Reger, B., Wootan, M.G. and Booth-Butterfield, S. 1999. Using Mass Media to Promote Healthy Eating: A Community-Based Demonstration Project. *Preventive Medicine.* **29**(5), pp.414-421.

Reger, B., Wootan, M.G. and Booth-Butterfield, S. 2000. A comparison of different approaches to promote community-wide dietary change. *American journal of preventive medicine.* **18**(4), pp.271-275.

Reger, B., Wootan, M.G., Booth-Butterfield, S. and Smith, H. 1998. 1% or less: a community-based nutrition campaign. *Public Health Reports.* **113**(5), pp.410-419.

Revoredo-Giha, C., Lamprinopoulou, C., Toma, L., Leat, P.M.K., Kupiec-Teahan, B. and Cacciolatti, L. 2009. Bread prices, consumption and nutrition implications for Scotland: a regional analysis using supermarket scanner data. *Land Economy Working Paper Series - Scottish Agricultural College.* (48), pp.23-pp.

Scarborough, P., Rayner, M., Harrington, R. 2021. *foodDB.* [Online]. Available from: https://www.ndph.ox.ac.uk/food-ncd/archive/research-projects/fooddb-and-myshop

Schwartz, J., Mochon, D., Wyper, L., Maroba, J., Patel, D. and Ariely, D. 2014. Healthier by precommitment. *Psychological science.* **25**(2), pp.538-546.

Serra-Majem, L., MacLean, D., Ribas, L., Brulé, D., Sekula, W., Prattala, R., Garcia-Closas, R., Yngve, A., Lalonde, M. and Petrasovits, A. 2003. Comparative analysis of nutrition data from national, household, and individual levels: results from a WHO-CINDI collaborative project in Canada, Finland, Poland, and Spain. *Journal of Epidemiology and Community Health.* **57**(1), pp.74-80.

Silver, L.D., Ng, S.W., Ryan-Ibarra, S., Taillie, L.S., Induni, M., Miles, D.R., Poti, J.M. and Popkin, B.M. 2017. Changes in prices, sales, consumer spending, and beverage consumption one year after a tax on sugar-sweetened beverages in Berkeley, California, US: A before-and-after study. *PLoS Medicine.* **14**(4), pe1002283.

Smith, D.M., Clarke, G.P., Ransley, J. and Cade, J. 2006. Food Access and Health : A Microsimulation Framework for Analysis. *地域学研究.* **35**(4), pp.909-927.

Stead, M., MacKintosh, A.M., Findlay, A., Sparks, L., Anderson, A.S., Barton, K. and Eadie, D. 2017. Impact of a targeted direct marketing price promotion intervention (Buywell) on food-purchasing behaviour by low income consumers: a randomised controlled trial. *Journal of human nutrition* *and dietetics : the official journal of the British Dietetic Association.* **30**(4), pp.524-533.

Sturm, R., Patel, D., Alexander, E. and Paramanund, J. 2016. Seasonal cycles in food purchases and changes in BMI among South Africans participating in a health promotion programme. *Public Health Nutrition.* **19**(15), pp.2838-2843.

Surkan, P.J., Tabrizi, M.J., Lee, R.M., Palmer, A.M. and Frick, K.D. 2016. Eat Right-Live Well! Supermarket intervention impact on sales of healthy foods in a low-income neighborhood. *Journal of Nutrition Education and Behavior.* **48**(2), pp.112-121.

Taylor, A., Wilson, F., Hendrie, G.A., Allman-Farinelli, M. and Noakes, M. 2015. Feasibility of a Healthy Trolley Index to assess dietary quality of the household food supply. *British Journal of Nutrition.* **114**(12), pp.2129-2137.

Theis, D.R.Z. and White, M. 2021. Is Obesity Policy in England Fit for Purpose? Analysis of Government Strategies and Policies, 1992–2020. *The Milbank Quarterly.* **99**(1), pp.126-170.

Timmins, K.A., Green, M.A., Radley, D., Morris, M.A. and Pearce, J. 2018. How has big data contributed to obesity research? A review of the literature. *International Journal of Obesity.* **42**(12), pp.1951-1962.

Tin, S.T., Ni Mhurchu, C. and Bullen, C. 2007. Supermarket sales data: Feasibility and applicability in population food and nutrition monitoring. *Nutrition Reviews.* **65**(1), pp.20-30.

Toft, U., Winkler, L.L., Mikkelsen, B.E., Bloch, P. and Glumer, C. 2017. Discounts on fruit and vegetables combined with a space management intervention increased sales in supermarkets. *European journal of clinical nutrition.* **71**(4), pp.476-480.

Tran, L.T.T., Brewster, P.J., Chidambaram, V. and Hurdle, J.F. 2017. An innovative method for monitoring food quality and the healthfulness of consumers' grocery purchases. *Nutrients.* **9**(5), p457.

Uusitalo, L., Erkkola, M., Lintonen, T., Rahkonen, O. and Nevalainen, J. 2019. Alcohol expenditure in grocery stores and their associations with tobacco and food expenditures. *BMC Public Health.* **19**(787).

Van Gestel, L., Kroese, F. and De Ridder, D. 2018. Nudging at the checkout counter-A longitudinal study of the effect of a food repositioning nudge on healthy food choice. *Psychology & Health.* **33**(6), pp.800-809.

Vandenbroele, J., Slabbinck, H., Kerckhove, A.v. and Vermeir, I. 2018. Curbing portion size effects by adding smaller portions at the point of purchase. *Food Quality and Preference.* **64**, pp.82-87.

Walmsley, R., Jenkinson, D., Saunders, I., Howard, T. and Oyebode, O. 2018. Choice architecture modifies fruit and vegetable purchasing in a university campus grocery store: time series modelling of a natural experiment. *BMC Public Health.* **18**(1), p1149.

Wilkins, E., Radley, D., Morris, M., Hobbs, M., Christensen, A., Marwa, W.L., Morrin, A. and Griffiths, C. 2019. A systematic review employing the GeoFERN framework to examine methods, reporting quality and associations between the retail food environment and obesity. *Health Place.* **57**, pp.186-199.

Chapter 4
Exploring the geographic variation in fruit and vegetable purchasing behaviour using supermarket transaction data

Victoria Jenneson, Graham P Clarke, Darren C Greenwood, Becky Shute, Bethan Tempest, Tim Rains, Michelle A Morris

Nutrients. **2022.** *14*(1), 177:, https://doi.org/10.3390/nu14010177

Keywords: supermarket, transaction data, food purchase behaviour, fruit and vegetables, dietary assessment, geographic clustering
Issue Section: Nutrition Methodology & Assessment

Abstract

The existence of dietary inequalities is well-known. Dietary behaviours are impacted by the food environment and are thus likely to follow a spatial pattern. Using 12 months of transaction records for around 50,000 ‘primary’ supermarket loyalty card holders, this study explores fruit and vegetable purchasing at the neighbourhood level across the city of Leeds, England. Determinants of small-area-level fruit and vegetable purchasing were identified using multiple linear regression. Results show that fruit and vegetable purchasing is spatially clustered. Areas purchasing fewer fruit and vegetable portions typically had younger residents, were less affluent, and spent less per month with the retailer.

4.1 Introduction

Poor dietary quality contributes to rising rates of obesity and associated comorbidities in the UK (NHS, 2019; NHS Digital., 2019). Many years of policies to encourage individual behaviour change have done little to reverse obesity rates (Theis and White, 2021). Moreover, the influence of the food environment on obesity and poor diets (Foresight, 2007; Nestle and Jacobson, 2000) has attracted policy attention (Ogden et al., 2001). Measures such as changes to food promotions (DHSC, 2019; DHSC, 2020) and the soft drinks industry levy (HMRC, 2018) in the UK have focused on altering the food environment to ‘nudge’ people towards healthier choices. The food industry has also taken voluntary action to make healthier diets more achievable, such as committing to selling more portions of vegetables as part of the Peas Please campaign (Food Foundation., 2021; Food Foundation., 2020a).

Studies of dietary behaviours are important for monitoring population dietary trends and responses to interventions such as policy changes. Population dietary assessment typically employs national survey data, such as the UK's National Diet and Nutrition Survey (NDNS) (GOV.UK, 2016). Surveys employ self-report methods such as food diaries and food frequency questionnaires, and offer detailed information on diet and nutrition as well as participant characteristics. This makes them useful for understanding the socio-demographic determinants of diet (Adams et al., 2015; Gibson and Neate, 2007; Maguire and Monsivais, 2015; Yau et al., 2019). However, the time and cost burdens for participants to complete surveys, and for researchers to code their outputs, limits their sample sizes. Relatively low sample sizes mean that the spatial resolution of national surveys is often poor and rarely offers detail below the regional level; regions in England have an average population greater than 5 million (Scarborough, 2008). This limits their utility to investigate spatial dietary inequalities which often occur at the neighbourhood level.

These surveys enable us to monitor and understand consumption of fruits and vegetables which in turn can be used as a proxy for a healthy diet due to their role in prevention of non-communicable diseases like cancer (WCRF, 2018), due to their richness in beneficial micronutrients, fibre and non-nutritive compounds, and their low energy density. Average fruit and vegetable consumption in the UK is below recommended levels in all ages (Public Health England., 2019), particularly among low-income groups (Public Health England., 2019). Existing dietary inequalities, especially in vegetable intake (Food Foundation., 2020b), have further deepened as a result of the COVID-19 pandemic (National Food Strategy., 2020), highlighting the need for additional action.

The inequalities in non-communicable disease rates and life expectancy seen at the neighbourhood level (McCartney, 2011) suggest that diets may follow spatial patterns similar to those observed for deprivation. This is supported by the food environment literature which considers access to 'healthy' and 'unhealthy' food outlets. Deprived areas are more likely to display a disproportional density of fast-food outlets (Fraser et al., 2010), and convenience shops, which are less and lack variety in their fruit and vegetable offering (Blake, 2019). Spatial exploration of supply-side characteristics, such as food environment exposures, is important for revealing inequalities in exposures and have led to planning policies banning fast-food outlets near schools by some local authorities in England (PHE, 2014).

However, previous studies have found the relationship between accessibility to food environment exposures and diet and health outcomes to be non-stationary over space (Fraser et al., 2012; Clary et al., 2016), suggesting moderation by uncaptured environmental and/or social determinants. As neighbourhood food availability does not necessarily translate to dietary behaviours among the individuals and households who live there, there is a need for large-scale exploration of demand-side diet-related behaviours at the small-area level, which has been lacking previously due to the limited spatial scale afforded by dietary survey data. That said, there is some evidence from survey data that shows that dietary quality varies spatially in line with the socioeconomic gradient. Healthier diets and higher fruit and vegetable intakes were found in neighbourhoods with a higher socioeconomic status in several countries (Menezes et al., 2017; Ball et al., 2015; Drewnowski et al., 2016; Morris et al., 2016). However, it is not easy to examine diet at the small-area level using traditional dietary assessment approaches, without which much of the local nuance is likely to be missed.

Considering purchases as an upstream behaviour for consumption, supermarket transaction records have been proposed as complementary to dietary surveys (Green et al., 2020), with the capacity to provide additional insight as a result of their granularity. Automatically generated electronic food purchase data have the potential to offer large volumes of geocoded information about household food and nutrition availability (Hamilton et al., 2007; Aiello et al., 2020; Närhinen et al., 1999; Clark et al., 2021). Using transaction records for loyalty card holders at a UK supermarket chain, this paper explores small-area and demographic variations in fruit and vegetable purchases (including fresh, frozen and dried varieties) that exist within a single city (Leeds, England). Additionally, we identify determinants of neighbourhood fruit and vegetable purchase levels. Given that areas of similar sociodemographic profile tend to cluster together, we anticipate a spatial patterning of fruit and vegetable purchases. This paper offers a novel exploration of the small-area geography of actual dietary purchase behaviours, as opposed to exposure. Thus, we provide a step towards an incorporated study of both supply and demand, which is likely to provide greater insight into how people’s interactions with their food environments shape their dietary habits. Revealing area-level characteristics which put residents at risk for low purchasing of fruits and vegetables may be used to better understand drivers of diet-related health inequalities and to target local interventions.

This paper will:

1. Examine the small-area spatial distribution of fruit and vegetable purchases and predictors of this purchase behaviour
2. Explore associations at a neighbourhood level between mean daily fruit and vegetable portions purchased and area socioeconomic characteristics, customer demographics, and access to supermarkets.
3. Develop a statistical model that identifies drivers of fruit and vegetable purchasing at a neighbourhood level.

4.2 Methods

4.2.1 Study sample

The study sample included 50,917 customers who held a loyalty card for a major UK supermarket, registered to an address in the city of Leeds, England. Eligible customers made at least ten transactions during 2016, which included a minimum of seven out of 16 food categories, developed from categories captured by the Living Costs and Food Survey (LCFS) (ONS, 2017b) (Table 4.1). The inclusion criteria are described in more detail elsewhere (Clark et al., 2021), but briefly they aim to capture 'primary' shoppers who do the majority of their food shopping with the study retailer. The median shopping frequency of our sample is 53 occasions annually (interquartile range 33 – 82) (Clark et al., 2021). Thus, we exclude customers with infrequent purchases from a limited range of food categories, on the basis that their purchases are unlikely to represent their overall diet.

Exploratory data analysis identified some customers with extremely high loyalty card expenditure which we considered unlikely to represent typical household purchasing. We defined an upper bound of annual expenditure, based on household expenditure on food and non-alcoholic beverages from the 2016 edition of the Family Food Survey (FFS) (ONS, 2017b)). A threshold of 1.5 times the inter-quartile range beyond the upper quartile (a common criteria to identify large outliers in box plots) from the FFS report, was used to exclude customers at the upper end of the expenditure distribution. For symmetry, the same proportion of customers (1.95%) at the bottom end of the annual expenditure distribution was removed. Customers must be aged 18 or over to obtain a loyalty card with the retailer. For this reason, we excluded customers with a recorded age of 17 years or below as these were assumed to be data errors. Anonymised customer characteristics (age, gender, and output area of residence) were derived from the retailer's loyalty card sign-up questionnaire. We assume that the loyalty card holder is the main person responsible for shopping in the household.

Table 4.1 Food categories in the transaction database used for sampling

[1]Categories based on the Living Costs of Food Survey Categories

Category description[1]	**LCFS category (LCFS code)**
Carbohydrate products	Bread and cereals (1.1.1)
Cakes and biscuits	Buns, cakes, biscuits etc (1.1.3)
Meat and fish	Meat (1.1.5 – 1.1.10), Fish (1.1.11)
Dairy	Milk, cheese, eggs (1.1.12 – 1.1.15)
Fats	Oils and fats (1.1.16 – 1.1.18)
Fruit	Fruit (1.1.19 – 1.1.22)
Vegetables and salad	Vegetables (1.1.23 – 1.1.27)
Potato	Potatoes (1.1.26)
Sweets	Sugar, jam, honey, chocolate confectionary (1.1.28 – 1.1.32)
Other (e.g. spices)	Other foods (1.1.33)
Non-alcoholic beverages	Non-alcoholic beverages (1.2)
Alcoholic beverages	Alcoholic beverages (2.1)
Ready foods	N/A – additional category not present in the LCFS
Baby food	N/A – additional category not present in the LCFS
Crisps and nuts	N/A – additional category not present in the LCFS
Meat free and free from foods	N/A – additional category not present in the LCFS

4.2.2 Study region

The study region is determined by customers whose loyalty card is registered to an output area inside the Leeds boundary. Leeds is a diverse city with cosmopolitan (ethnically diverse) and deprived areas in the south and west of the city, affluent suburbs in the north and east, and a large student population in the inner western suburbs (Figure 4.1). Figure 4.1a shows the spatial distribution of the 2015 Index of Multiple Deprivation (IMD) decile at the Lower Super Output Area (LSOA) level, a neighbourhood census geography representing 400 – 1200 households. The IMD is a rank of deprivation for more than 32,000 LSOAs in England (GOV.UK, 2015). These are split into deciles, where 1 represents the most deprived 10% of areas in England. Figure 4.1b shows the 2011 UK Output Area Classification (OAC) for Output Areas (OAs) in Leeds; the OA is a small-area census geography containing around 125 households. The OAC is an open-source census-derived national hierarchical geodemographic classification system (Vickers et al., 2005; Gale et al., 2016).

Customer area of residence is known at the Output Area (OA) level and is used to describe the characteristics of areas in the study in the absence of detailed individual-level demographic data. This study uses the Supergroup level of the OAC hierarchy, which assigns areas to one of eight Supergroups, according to the affluence, ethnic composition, rurality, age demographics and other characteristics of the people residing there. Due to small customer numbers at the OA level, areas were aggregated to the Lower Super Output Area (LSOA) (400 – 1200 households) (ONS, 2017a) for analysis. LSOAs with low customer numbers (<n=10) were excluded. The OA of residence centroid was used to assign eastings and northings for customer residential location. This was used to calculate Euclidean (straight-line distance) distance to the nearest store and most frequently used store, from the study retailer.

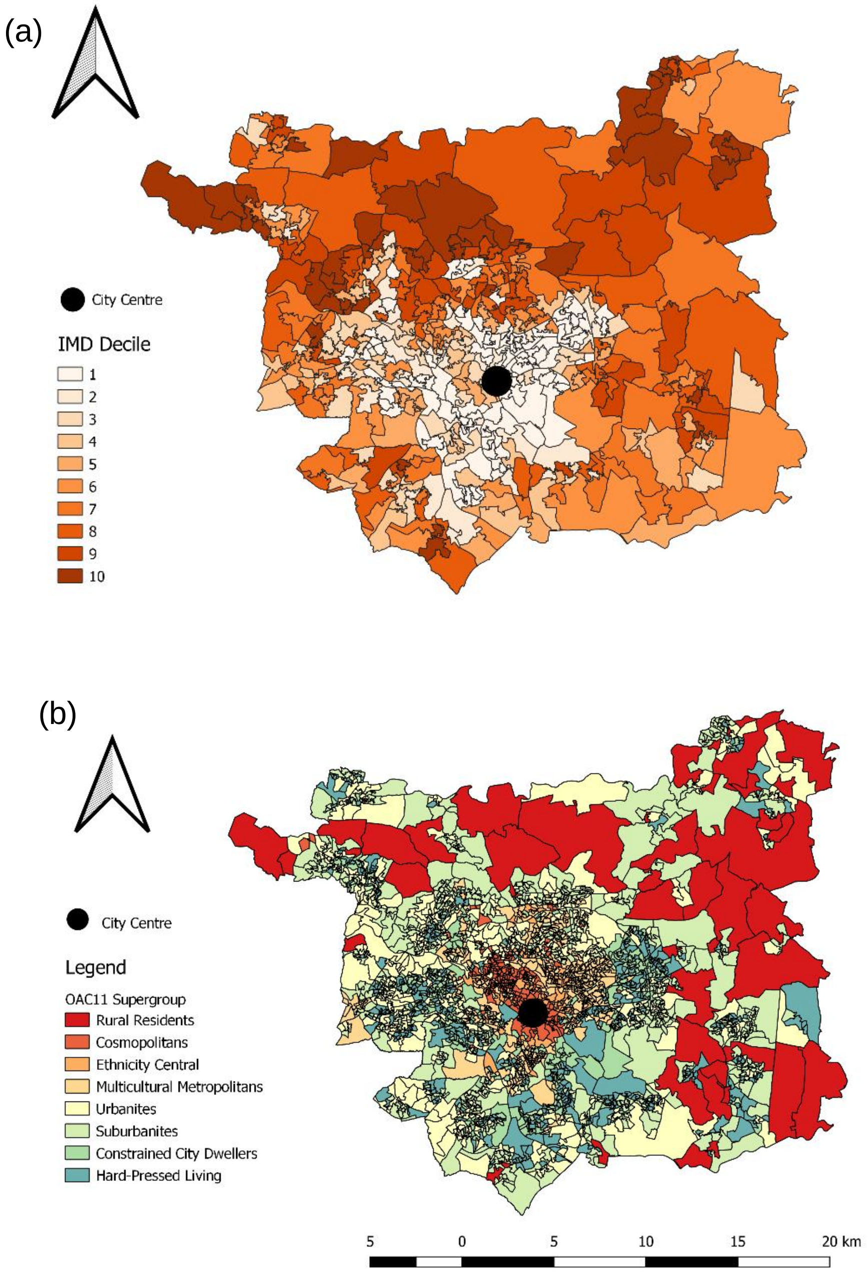


Figure 4.1 (a) Index of Multiple Deprivation decile by Lower Super Output Area in Leeds. (b) Output Area Classification by Output Area in Leeds.

4.2.3 Transaction data

All loyalty card transactions made with the retailer by our 'primary shopper cohort' (online and in any store regardless of format, including those made outside of the study region) were collected for the 2016 calendar year. Items in the transaction database have a corresponding weight and number of units, used to calculate their quantity by weight or volume. Non-food and beverage transactions were removed from the database by the retailer prior to access by the research team. Transactions are linked to the loyalty card holder by a unique hashed customer pseudo-ID, and items purchased on a single occasion are linked by a transaction ID.

4.2.4 Estimating fruit and vegetable purchases

Item sub-categories used by the retailer were mapped to categories from the LCFS (Table 4.1) (ONS, 2017b). Fruit and vegetable purchases were then identified by selecting the relevant LCFS categories (Fruit, Vegetables and salad Table 1). The LCFS is a granular database containing approximately 80 food categories, allowing for the exclusion of potatoes (in line with the UK's 5-a-day fruit and vegetable consumption guidance (NHS, 2018a)), and inclusion of both fresh and processed (e.g., frozen and canned) fruits and vegetables. As ready meals were coded as a separate category their constituent parts are not quantified. Therefore, any fruit or vegetables purchased as part of ready meals are not accounted for.

Mean daily fruit and vegetable portions purchased were calculated for each household, by dividing their total purchased weight (grams) in 2016 by 80 (the number of grams recommended as a portion of fresh fruits and vegetables in the UK's 5-a-day recommendation (NHS, 2018b)), and then further dividing by 366 (2016 was a leap year). While dried fruits and pulses contributed to the overall fruit and vegetable purchased weight, their recommended portion size (30g) was not explicitly accounted for, due to challenges in data format allowing accurate identification of them, underestimating their contribution to purchased portions. In the absence of supplementary survey data, food waste and the edible proportions of fruits and vegetables were not accounted for, nor was the number of people living in the household.

4.2.5 Analysis

This study used a multiple linear regression model to identify drivers of mean daily fruit and vegetable purchasing at the neighbourhood level. Model parameters were chosen to represent three domains which were considered to be theoretically influential for dietary choices; customer demographic characteristics (% females, and % aged 65+ years [other age groups were omitted due to lack of influence on the model]); neighbourhood characteristics (mean Index of Multiple Deprivation decile and % of customers in each OAC Supergroup); and accessibility metrics (mean distance to nearest store, mean distance to most-used store, and shopping frequency [mean monthly food and beverage transactions]). Mean total monthly spend on food and beverages (£) was also controlled for.

Outliers from the model were identified as those LSOAs with a Cooks Distance (accounting for leverage and residuals) greater than 0.009, using the threshold 4/n (n = 439). The model was then reapplied after exclusion of model outliers, which allowed for exploration of the characteristics of outlier neighbourhoods.

Prior to building the regression model, the correlation between mean daily fruit and vegetable portions purchased and each predictor variable was estimated using Kendall’s Tau correlation to inform variable selection for the regression model. Secondly, spatial autocorrelation of each variable was explored using the univariate Moran’s I (Index), to inform the need for a Geographically Weighted Regression model (GWR). LSOAs without any customers were omitted from the Moran’s I calculation. Moran’s I may hold a value from -1 (indicating perfect dispersion) to 1 (indicating perfect clustering), where 0 indicates random dispersion. For the purpose of this study, values smaller than -0.5 are considered as evidence of dispersion, while values greater than 0.5 are considered evidence of clustering, if they are significant at the 95% confidence level. Exploration of the explanatory variables revealed spatial clustering only in those variables which were inherently spatial in nature (IMD and OAC supergroup). For this reason, it was considered that their inclusion in an Ordinary Least Squares (OLS) regression model should be sufficient to capture much of the neighbourhood variation in the outcome, and a GWR model was not used.

4.3 Results

4.3.1 Customer characteristics

The data covers 50,917 loyalty card holders, equivalent to approximately 6% of the Leeds population. However, as loyalty cards typically represent a household, in reality our sample likely accounts for a larger proportion of residents. Without detailed household size information for the study sample, the exact number of people captured is unknown, but using the average household size for Leeds (2.3 people (Leeds Observatory., 2021)), we estimate it represents approximately 117,000 people (around 15% of the Leeds population).

A summary of customer characteristics, compared with demographics for Leeds overall, is shown in Table 4.2. The number of female loyalty card holders was more than double the number of male loyalty card holders. Almost 40% of customers are in the 45-64 age band, which is over-represented compared with the Leeds population. The sample over-indexes on customers living in affluent regions; more than 72% of customers live in LSOAs in the five most affluent deciles, compared with less than 43% of the general Leeds population. Compared with the population of Leeds, the customer sample over-indexes on customers from Rural Residents, Urbanites and Suburbanites, and under-represents people from areas classified as Cosmopolitans, Ethnicity Central, Multicultural Metropolitans, Constrained City Dwellers, and Hard-pressed Living supergroups.

4.3.2 Fruit and vegetable purchases in Leeds

In 2016, customers across Leeds purchased on average 3.4 portions (equivalent to 272 g) of fruit and vegetables per household per day (Table 2). This is equivalent to around 1.5 portions per person per day, given the average household size of 2.3 persons (Leeds Observatory., 2021). Mean fruit and vegetable portions when aggregated across LSOAs was lower at 3.0/household/day (Table 3), highlighting that accounting for local averages can mask local patterns. Female loyalty card holders purchased on average 0.23 portions more per day for their household than males. Younger adults purchased fewer daily portions per household of fruits and vegetables (mean = 2.96 per for 18 – 44 years) compared with older adults (mean = 3.64 for adults age 65+). Customers living in the most deprived areas (IMD decile 1, mean = 2.80 portions per household) purchased on average 1.12 portions per household of fruits and vegetables fewer each day compared with customers in the most affluent areas (IMD decile 10, mean = 3.92 portions per household). Customers living in Suburbanite areas had the highest purchases of fruits and vegetables (3.77 portions/household/day), while those in Cosmopolitan areas purchased the fewest portions (2.58 portions/household/day), a difference of 1.19 daily portions.

Table 4.2 Coverage of study sample by demographic group, in relation to Leeds and UK

[1]Leeds population figures (gender and age) from 2011 UK census, n = 751, 485 residents (Office for National Statistics., 2013). IMD data from 2015/16 by LSOA, n = 784,846 residents (Leeds Observatory., 2019). OAC Supergroup population estimates derived from 2016 mid-year population estimates (n = 781,087 residents) (Office for National Statistics., 2021). FV = Fruits and Vegetables.

		Number (%)		
Characteristic		**Study population**	**Leeds population[1]**	**Mean daily portions of FV purchased per household (SD)**
Whole sample		50,917 (100.0)	751,485 (100)	3.40 (3.06)
Gender	Male	14,539 (28.6)	367,933 (49.0)	3.22 (2.98)
	Female	32,342 (63.5)	383,550 (51.0)	3.45 (3.10)
	Unknown	4,036 (7.9)	-	3.69 (3.07)
Age band	18 - 44	16,268 (32.0)	269,582 (35.9)	2.96 (2.80)
	45 – 64	19,614 (38.5)	172,964 (23.0)	3.58 (3.27)
	65+	10,817 (21.2)	109,598 (14.6)	3.64(2.99)
	Unknown	4,218 (8.3)	-	3.65 (3.04)
IMD decile	1	3,621 (7.1)	186,995 (23.8)	2.80 (2.57)
	2	2,035 (4.0)	75,224 (9.6)	2.70 (2.47)
	3	2,669 (5.2)	70,571 (9.0)	2.77 (2.56)

	4	1,903 (3.7)	33,388 (4.3)	2.86 (2.81)
	5	3,769 (7.4)	83,694 (10.7)	2.92 (2.66)
	6	4,770 (9.4)	68,864 (8.8)	3.20 (3.00)
	7	7,650 (15.0)	89,670 (11.4)	3.48 (3.11)
	8	7,573 (14.9)	63,366 (8.1)	3.47 (3.10)
	9	8,974 (17.6)	62,882 (8.0)	3.84 (3.26)
	10	7,953 (15.6)	50,192 (6.4)	3.92 (3.33)
Output area Classification Supergroup	Rural Residents	1,428 (2.8)	12,844 (1.6)	3.58 (3.11)
	Cosmopolitans	3,839 (7.5)	80,788 (10.3)	2.58 (2.39)
	Ethnicity Central	731 (1.4)	28,615 (3.7)	2.88 (2.48)
	Multicultural Metropolitans	4,889 (9.6)	140,250 (18.0)	3.21 (2.98)
	Urbanites	14,784 (29.0)	161,993 (20.7)	3.50 (3.10)
	Suburbanites	18,445 (36.2)	160,366 (20.5)	3.77 (3.27)
	Constrained City Dwellers	1,949 (3.8)	71,244 (9.1)	2.67 (2.47)
	Hard-pressed Living	4,852 (9.5)	124,987 (16.0)	2.88 (2.72)

4.3.3 Neighbourhood characteristics

The characteristics of study areas (aggregated to the LSOA level) are summarised in Table 4.3. On average, customers in each LSOA live a median of 1.7 km from their nearest study retailer store (which may be a superstore or convenience format), but shop most often at stores further away (a median of 11.2 km away). The average spend with the retailer across LSOAs is £104 per month. Customers have a median shopping frequency with the retailer of just over 5 occasions each month, indicating relative loyalty to the study retailer.

The outcome variable, mean daily fruit and vegetable portions, shows evidence of spatial clustering (Moran's I 0.52, $p<0.001$). Evidence of spatial clustering was also found for IMD decile (Moran's I 0.61, $p<0.001$), % customers in the Cosmopolitan, Ethnicity Central and Multicultural Metropolitans supergroups (Moran's I = 0.71, 0.58 and 0.60 respectively, all $p<0.001$), and mean distance to nearest store (Moran's I 0.83, $p<0.001$), while mean total monthly spend was borderline (Moran's I 0.49, $p<0.001$). As no evidence of spatial clustering or dispersion was found for the other predictor variables, we accept the null hypothesis that their spatial distribution is random.

Mean total monthly spend (£) was found to be significantly correlated with the outcome (mean daily fruit and vegetable portions/household), indicating that as monthly expenditure increases so does the number of fruit and vegetable portions purchased ($C = 0.7$, $p<0.001$). The correlation between IMD decile and mean daily fruit and vegetable portions was also positive and reached statistical significance at the 95% level, though moderate in strength ($C = 0.5$, $p<0.001$), indicating that as affluence increases the number of fruit and vegetable portions purchased increases.

Table 4.3 Overview of variables at Lower Super Output Area level

For variables which did not display a normal distribution, the median and interquartile range (IQR) are the provided summary statistics. FV = Fruits and Vegetables

Characteristic of loyalty card holder	Mean (SD) [1]Median (IQR)	Univariate Moran's I (clustering)	p-value (Moran's I)	Kendall's Tau rank correlation with outcome	p-value (Kendall's Tau)
Outcome variable					
Mean household daily portions of FV purchased	3.0 (0.7)	0.5	0.001	-	-
Predictor variable					
female (% of sample)	63.6 (8.1)	0.1	0.006	0.0	0.515
aged 18 – 44 years (% of sample)	34.3 (15.3)	0.4	0.001	-0.3	<0.001
% aged 45 – 64 years (% of sample)	38.6 (9.8)	0.2	0.001	0.1	0.002

% aged 65+ years (% of sample)	19.1 (9.8)	0.3	0.001	0.3	<0.001
IMD decile	5.2 (3.1)	0.6	0.001	0.5	<0.001
Rural Residents (% of sample)	0.0 (0.0, 0.0)[1]	0.3	0.001	0.2	<0.001
Cosmopolitans (% of sample)	0.0 (0.0, 0.0)[1]	0.7	0.001	-0.1	0.066
Ethnicity Central (% of sample)	0.0 (0.0, 0.0) [1]	0.6	0.001	-0.1	<0.001
Multicultural Metropolitans (% of sample)	0.0 (0.0, 20.3) [1]	0.6	0.001	-0.1	<0.001
Urbanites (% of sample)	0.0 (0.0, 41.6) [1]	0.3	0.001	0.2	<0.001
Suburbanites (% of sample)	0.0 (0.0, 45.3) [1]	0.4	0.001	0.4	<0.001
Constrained City Dwellers (% of sample)	0.0 (0.0, 7.1) [1]	0.2	0.001	-0.3	<0.001
Hard-pressed Living (% of sample)	0.0 (0.0, 24.4) [1]	0.2	0.001	-0.2	<0.001

Mean distance to nearest store (km)	1.7 (0.9, 2.8)[1]	0.8	0.001	0.1	<0.001
Mean distance to most used store (km)	11.2 (6.4, 17.6)[1]	0.4	0.001	-0.1	<0.001
Mean total monthly spend (£)	104.3 (19.1)	0.5	0.001	0.7	<0.001
Shopping frequency (mean monthly trips)	5.0 (4.4, 6.0)[1]	0.4	0.001	-0.1	<0.001

4.3.4 Spatial patterns in fruit and vegetable purchasing

Figure 4.2 shows the spatial pattern of fruit and vegetable purchases across Leeds at the LSOA-level. Fruit and vegetable purchasing is spatially clustered (Figure 4.3) and follows the expected deprivation trend, with the most deprived purchasing fewer fruit and vegetable portions. Households living in the North of Leeds purchase on average 4 or more portions/day of fruit and vegetables. The more multicultural and urban areas in the centre and South-West of Leeds purchase the fewest daily fruit and vegetable portions. Those more rural and suburban areas surrounding the city centre, particularly to the North and East, purchase 3 – 4 portions/day on average per household.

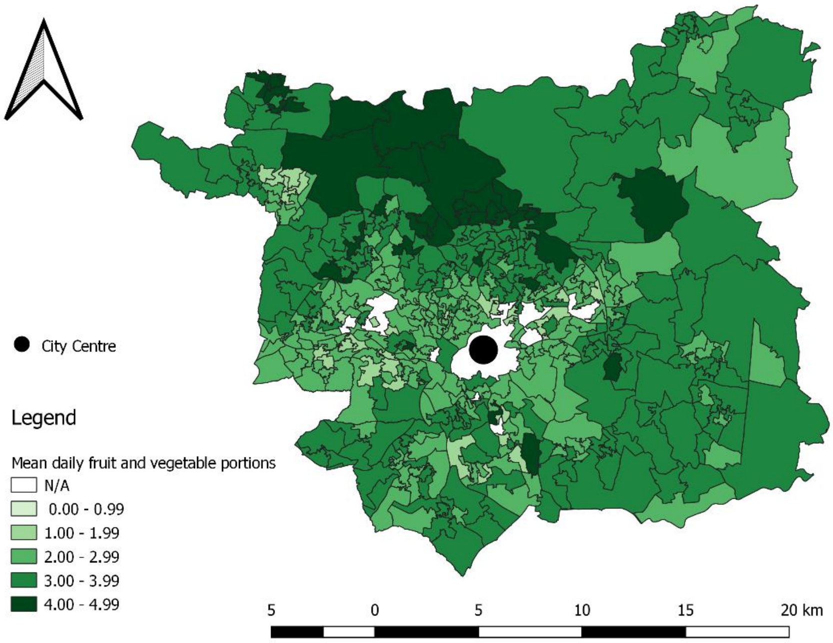


Figure 4.2 Fruit and vegetable purchasing in Leeds by Lower Super Output Area: mean daily portion per household.

Areas with N<10 customers omitted from map (shown as N/A in the figure legend)

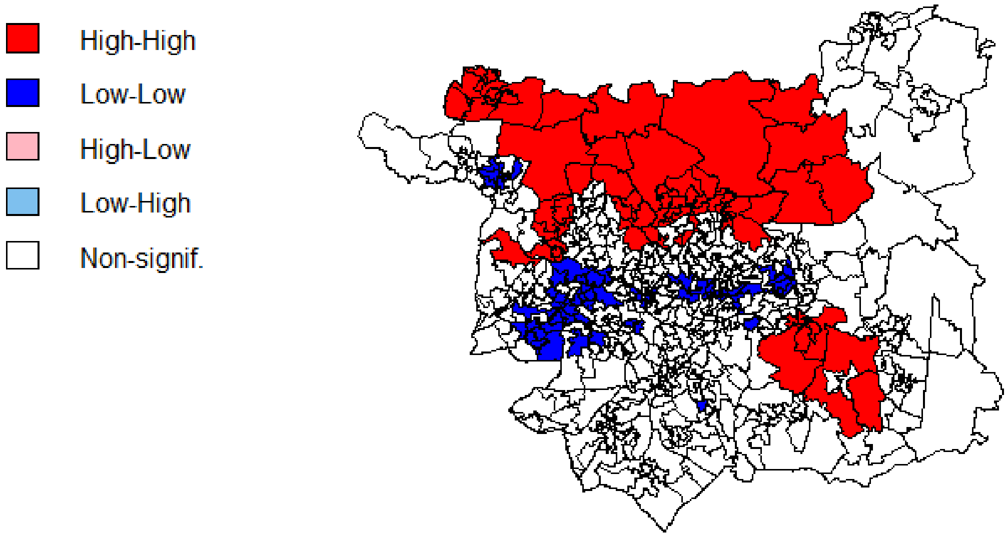


Figure 4.3 Local Moran’s I for daily fruit and vegetable portions per household

4.3.5 Linear regression

Regression coefficients for all LSOAs (n=439) are shown in Table 4.4. Lower deprivation and a greater proportion of older adults (65+ years) are positively associated with mean daily fruit and vegetable portions per household purchased at the LSOA level. Mean daily fruit and vegetable purchases among LSOAs in IMD decile 10 (the least deprived) are around 0.45 portions per household higher than the most deprived LSOAs (IMD decile 1). Theoretically, an area where 100% of the population are aged 65 years or older is likely to purchase half a portion/household/day more fruits and vegetables on average than an area where only 1% of the population are aged 65+. The proportion of female customers did not affect household fruit and vegetable purchasing at the LSOA-level, but it was influenced by output area classification. A higher proportion of customers living in neighbourhoods classified as Cosmopolitans, Ethnicity Central, Multicultural Metropolitan and Suburbanites, was significantly associated with higher fruit and vegetable purchases, while Constrained City Dwellers were associated with fewer fruit and vegetable portions. A higher mean total monthly expenditure with the retailer was associated with a greater number of fruit and vegetable portions purchased. A £1 increase in LSOA-level mean total food and non-alcoholic beverage spend with the retailer was associated with an additional 0.03 portions fruits and vegetables/household/day purchased. Shopping frequency and distance to store were not associated with fruit and vegetable purchase levels.

25 outlier LSOAs were identified by the model and are summarised in Supplementary Table S1 and mapped in Supplementary Figure S2. Overall, outlier areas had a higher proportion of customers in the most deprived IMD decile, and decile 4, and a lower proportion of customers in the least deprived deciles, compared with the overall sample. These areas included deprived areas with higher fruit and vegetable purchases than expected and low deprivation areas with lower fruit and vegetable purchases than expected. Examination of the group and supergroup levels of the OAC classification also revealed that outlier areas were also more likely to be resided by ethnic minority communities.

Table 4.4 Results of OLS regression predicting household fruit and vegetable purchasing (portions/day)

[1]%OAC8 (Hard-pressed living) was excluded from the model due to perfect multicollinearity with the intercept

	OLS regression, n=439 LSOAs (Adj R2: 85.8%)	
Variable[1]	Coefficient (95% CI)	P-value
Intercept	-0.565 (-0.918, -0.213)	0.003
Mean monthly spend (£)	0.031 (0.029, 0.032)	<0.001
% aged 65+ years	0.005 (0.002, 0.008)	0.002
IMD decile	0.045 (0.028, 0.061)	<0.001
Shopping frequency (mean monthly trips)	0.026 (-0.001, 0.053)	0.066
% female	-0.003 (-0.007, -0.000)	0.057
Distance to nearest store (km)	0.006 (-0.021, 0.033)	0.654

Distance to most-used store (km)	0.001 (-0.001, 0.003)	0.280
% Rural Residents	-0.003 (-0.006, 0.001)	0.126
% Cosmopolitans	0.003 (0.001, 0.005)	0.011
% Ethnicity Central	0.004 (0.001, 0.007)	0.005
% Multicultural Metropolitans	0.002 (0.001, 0.003)	0.003
% Urbanites	0.002 (0.001, 0.003)	0.008
% Suburbanites	0.001 (-0.001, 0.002)	0.436
% Constrained City Dwellers	-0.002 (-0.003, 0.000)	0.093

LSOAs with high positive residual values (Figure 4.4) (≥0.5), indicating that customers in these areas purchase upwards of 0.5 portions more than predicted, tended to be dominated by OAC sub-groups characterised by families and ethnic minority groups. While those with high negative residuals (≤-0.5), indicating they purchase at least 0.5 portions fewer than predicted, tended to be dominated by OAC sub-groups characterised by retirement living or students, or families with a below average spend.

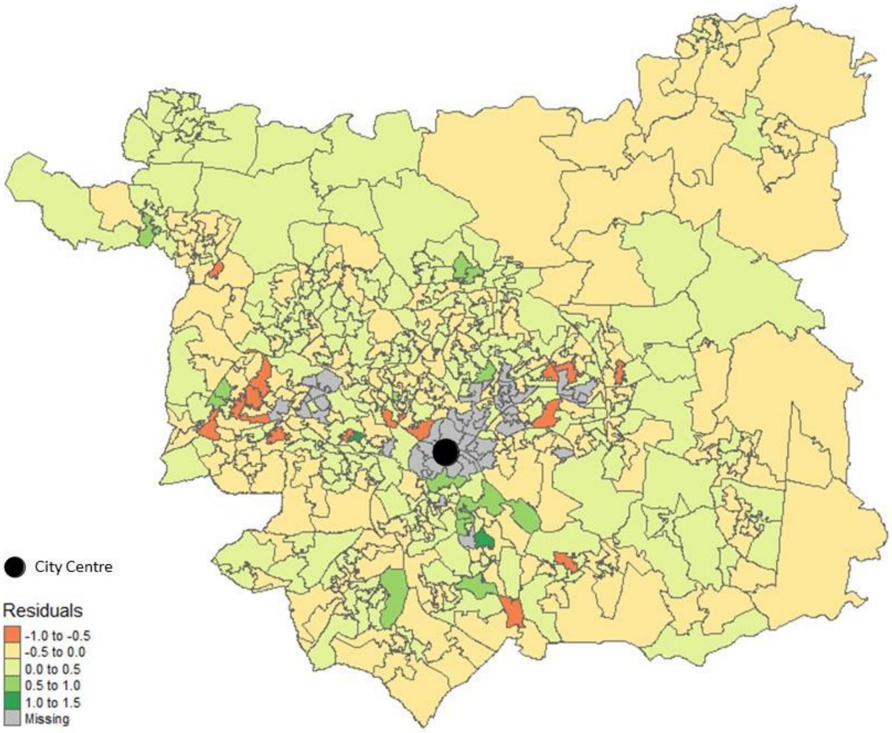


Figure 4.4 Map of residuals from linear regression model

4.4 Discussion

To the best of our knowledge, this is the first study to examine neighbourhood spatial variation in food purchases using electronic supermarket transaction records. Additionally, the ability to explore diet-related behaviours at such a fine geographic scale is a novel characteristic of purchase records. This study has several strengths including; the large sample size which affords statistical confidence in the results; geocoded dietary purchase data permitting visualisation and data linkage at the small-area level; objective dietary purchase estimates free from subject reporting biases and; longitudinal dietary purchase data for a whole year representing habitual dietary behaviours. Our findings demonstrate how novel exploration of large-scale purchase records at the neighbourhood geography level can offer an economical approach to population-level dietary assessment. Detecting socio-spatial influencers of dietary behaviours contributes to knowledge of localised dietary inequalities which are important for identifying potential intervention target areas.

Demographic information is available for this study thanks to loyalty card information provided by the retailer and linkage with area-level demographic data. This enables assessment of sample representativeness, which is noted as important (Rains and Longley, 2021) and lacking (Jenneson et al., 2021) in previous applications of transaction data for public health nutrition research. The customer sample are mostly female, with an older age distribution than Leeds as a whole. Affluent urban and suburban communities are over-represented while ethnically diverse communities are under-represented. Loyalty card customers introduce sampling bias, yet as a major cohort of the customer base, they make a useful research population. Despite the myth, surveys are not always more representative and tend to under-represent hard-to-reach low-income groups, especially those that use random sampling (Bonevski et al., 2014; Rehm et al., 2021). While some small geographic areas in this study have low customer numbers, the overall sample (n > 50,000) is very large compared with many presented in the literature and all socio-economic and geodemographic groups are represented in relatively large numbers (the lowest being 731 customers in the Ethnicity Central Output Area Classification Supergroup). Supermarket data, even from a single retailer, may therefore contain higher numbers of the hard-to-reach groups, giving greater power across all socioeconomic segments of the population. That said, we cannot be sure that customers in our sample are typical of their neighbourhood characteristics.

Customers in Leeds purchased on average 3.4 portions of fruits and vegetables per household per day, which equates to just 1.5 daily fruit and vegetable portions per person, considering the size of the average Leeds household (2.3 people) (Leeds Observatory., 2021). Our purchase estimate is well below the 5-a-day recommendation and lower than daily intakes estimated by the NDNS (4.2 portions per person) (Public Health England., 2016) and the Health Survey for England (HSE) (3.8 portions per person) (Osbourne et al., 2018). Survey estimates are known for over-reporting of fruit and vegetables due to social desirability biases, which are not a problem for objective automated purchase records.

The degree to which household-level purchases from the retailer represent individual consumption is unknown. Previous validation studies highlight that agreement between purchases and consumption is likely to vary by loyalty status and household composition (Vepsalainen et al., 2021; Eyles, H. et al., 2010), with higher agreement observed for single-person households (Vepsalainen et al., 2021). However, accepted adjustment factors remain lacking. Future work could incorporate known dietary variation by gender and life-stage by accounting for household composition (number and age of household members) to more accurately estimate individual-level intake from household purchase records. As this information cannot typically be obtained from retailer loyalty card records, this may involve using survey data, area-level estimates, or the development of methodologies to model household composition, for example microsimulation using census statistics (Robards et al., 2017; Birkin et al., 2018).

As we do not account for household waste or inedible proportions, our portions estimate may be inflated by as much as 28% for fresh vegetables and salad, and 6% for fresh fruit, according to national household waste estimates (Wrap, 2018). While robust methods for adjusting transaction records for waste are needed, crude application of national estimates would reduce our portions estimate to roughly 1.1 portions purchased per person per day. Furthermore, as our estimate is from a single retailer only, and does not include fruit and vegetables purchased or obtained elsewhere (e.g., from other retailers, home-grown, or consumed in restaurants) or in composite dishes purchased from the retailer, it is likely to under-represent total household fruit and vegetable purchases.

Fruit and vegetable purchases were found to vary spatially, with clusters of high fruit and vegetable purchasing in the affluent rural and suburban areas to the north and east of the city, while clusters of low fruit and vegetable purchasing were observed in the more deprived neighbourhoods in and around the city centre. The observed association between fruit and vegetable purchasing and area deprivation concurs with research into the geography of dietary patterns based on survey data, which found a higher prevalence of the vegetable-rich ‘health conscious’ and ‘high diversity vegetarian’ dietary patterns in suburban areas with lower deprivation (Morris, M.A. et al., 2014; Morris, M. et al., 2016). Using transaction records, fruit and vegetable purchases were important determinants of the observed ‘Fruity’ and ‘Meat Alternative’ dietary patterns, which were more prevalent among customers in the most affluent deciles (Clark et al., 2021). Yet, it is possible that the observed deprivation pattern may be confounded by differences in household composition, for example the mix of adults and children.

Despite the apparent presence of an overall deprivation gradient in fruit and vegetable choice behaviours, exploration of LOSAs classed as outliers and with high residual values identified neighbourhoods which appear to be exceptions to the rule. These areas suggest that education and ethnicity moderate the effect of deprivation. In spite of relative deprivation and a low overall spend, outlier areas occupied by students and minority ethnic families spent a higher-than-average proportion of their total expenditure on fruits and vegetables, which translated to more portions purchased than predicted. This could be indicative of a preference for scratch-cooking or meal assembly (e.g. the addition of peppers to a fajita meal kit) among these groups. Similarly, deprivation did not translate to low fruit and vegetable purchases for some rural communities. A higher than average spend observed in these outlier areas could be attributed to transactions capturing a larger proportion of total purchases, due to less retail competition. Despite spending a lower proportion of their total expenditure on fruits and vegetables, this did not translate to fewer portions, which may indicate thriftiness and a preference for cheaper fruit and vegetable varieties, which enable them to get more portions for their money.

Outlier LSOAs with lower than predicted fruit and vegetable purchases were occupied by families right across the deprivation spectrum. While these areas had a higher than average spend with the retailer, they prioritised spend on fruits and vegetables to a lesser degree. This may be indicative of busy family lives and a preference for convenience meals, a tendency to source fruits and vegetables elsewhere e.g. greengrocers or home-growing, or a preference for more expensive varieties. Outlier LSOAs also had a lower proportion of female customers overall, especially among more deprived areas. A sensitivity analysis repeating the model after exclusion of outlier LSOAs led to the proportion of females becoming a significant negative predictor of fruit and vegetable purchases (Supplementary Table S3). This is surprising given that females purchase more fruit and vegetables than males on average at the customer-level. While the reason is unclear, it could be that females are more likely to be the primary shopper for busy families which rely on convenience meals.

At the neighbourhood level, a higher proportion of over 65s was associated with higher fruit and vegetable portions purchased. The relationship with age may be a true reflection of differences in fruit and vegetable intake and agrees with other studies which found higher fruit and vegetable consumption among older adults (Aggarwal et al., 2014; NDNS, 2018; Public Health England., 2019). Yet, at the household level it is perhaps counter-intuitive that older adults should purchase more portions of fruit and vegetables, given that they are more likely to live alone or with just one other as children have left home. It is possible therefore that the relationship may also reflect differences in purchasing and food preparation practices. For example, younger adults often lack cooking skills, are likely to be under greater time-pressures due to work and childcare responsibilities, and may therefore prefer to choose convenience meals rather than cooking from scratch (Winkler and Turrell, 2010; Mills, 2018). While estimates by the retailer indicate that ready meals contribute only a small fraction of all vegetables purchased (unpublished data), our inability to accurately quantify the fruit and vegetable content of composite foods is likely to under-estimate fruit and vegetable purchases particularly among low-income working families and young people. Younger adults also consume more takeaway and restaurant meals (Adams et al., 2015), which may provide additional uncaptured fruit and vegetable portions.

Some research suggests that greater access to supermarkets is associated with higher fruit and vegetable intake (Menezes et al., 2017; Clary et al., 2016). Despite this, distance to nearest store and most used store were not found to be significantly associated with fruit and vegetable purchases in either model in this study. Indeed, rural and suburban areas to the north of the city demonstrated both the greatest average distances to nearest store and the highest fruit and vegetable purchases. It is possible that the relationship between proximity and fruit and vegetable purchases may vary spatially, moderated by unmeasured structural factors such as car ownership, access to public transport, store format (superstore or convenience store), the availability of other food outlets in the neighbourhood, and the degree to which a particular retailer meets a customer's social, cultural and economic needs (Clary et al., 2016). While all store formats offer some fruits and vegetables, there will be differences in the range offered. Aggarwal *et al* (2014) found that only one third of participants shopped at their nearest store, and those who shopped at low-cost stores were more likely to travel beyond their nearest store.

In another study by Liese *et al* (2014), access to store was associated with frequency of shopping trips, but not with fruit and vegetable intake, suggesting that access may be more closely associated with purchase pattern (e.g. top up shopping compared with a large weekly shop) than purchased amounts. While shopping frequency was not found to be significantly associated with fruit and vegetable purchases in the present study, we observed a narrowing of confidence intervals around our estimates after removal of outlier LSOAs, increasing the significance of findings (supported by a smaller p-value). Outlier areas were on average further from their most used store than the sample as a whole. The validity of distance as a measure of access should also be considered as it disregards the store offering and product prices. The average distance to the most-used store was high in this study (>10km), with a number of customers frequenting stores outside of the Leeds study region. While these are likely to be edge cases led by store network accessibility, this behaviour warrants further exploration. The high distance to most used store may be explained, for example, by customers shopping on their commute to work outside of the area, spending time at two addresses (for example students who return home outside of term time), or customers who have migrated outside the area without updating the address associated with their loyalty card.

The literature indicates good agreement between supermarket purchase data and self-reported dietary measures (Ransley et al., 2001; Eyles, Helen et al., 2010; Appelhans et al., 2017). Among loyal customers, even a single retailer can make a significant contribution to total household food purchases (Eyles, Helen et al., 2010; Hamilton et al., 2007; Hauser et al., 2013). While we do not know how much of a customer's total purchases are represented by the retailer, we have tried to select a relatively loyal customer sample, as indicated by their membership in the loyalty card scheme and frequent and broad-ranging purchase history. Customers in the sample visit the store on average 5 times per month. Controlling for total monthly spend on food and non-alcoholic beverages with the retailer goes some way to account for loyalty, assuming that higher spend with the retailer represents a higher proportion of the available food purse. Although, higher total monthly spend may also be indicative of a larger household size or affluence, denoting a preference for more expensive premium food stuffs rather than volume of food purchased. Degree of loyalty could better be controlled for using estimates of basket share or the Recency, Frequency, and Monetary value (RFM) index for example. Alternatively, as proposed by Rains and Longley (Rains and Longley, 2021), purchase 'completeness' at the category level could be estimated by comparing retail expenditure with estimates in national survey data.

While we observed spatial clustering of the outcome variable, the only predictor variables which showed spatial clustering were IMD and OAC, which are inherently spatial. As the deprivation index and geodemographic segmentation to go some way to capturing the nature of the food environment and the characteristics of people who live in an area, we considered the effect of uncaptured spatial factors on the model coefficients to be minimal. Despite this, we found LSOAs with high positive residual values to be clustered in the south of the city and those with high negative residual values to be clustered in the west. Similarly, Clary *et al* (Clary et al., 2016) found nonstationarity in the interaction between food environmental exposures and fruit and vegetable intake using GWR across four London boroughs. While there are likely to be limits to the validity of GWR at such granular geographic scales as that applied in this study, it is possible that our global model may have missed spatial variation in the local food environments and the way in which people respond to their environment. Incomplete spatial representation of dietary behaviours due to missing information about transactions from other retailers further limits the applicability of GWR approaches. Nevertheless, exploration of outlier areas from the regression model revealed some interesting insights which became more apparent when applying more granular levels of the hierarchical Output Area Classification (Group and Sub-group, rather than Supergroup as used in the model).

4.4.1 Policy relevance

Dietary research has long shown socioeconomic inequalities. While low overall fruit and vegetable purchase level warrant efforts to increase purchasing across the board, geographically untargeted strategies require huge investment and are likely to widen inequalities. To ensure those who purchase the least fruits and vegetables are not left behind, it is important to understand where best to focus interventions. Exploring neighbourhood-level fruit and vegetable purchases offers retailers insights for store-level stocking and marketing decisions. Interventions to increase fruit and vegetable purchases should target stores in areas with low purchase levels, especially those serving younger more deprived urban communities. These areas tend to be served by smaller stores where limited ranges make groceries comparatively more expensive. With small stores set to be exempt from new location-based in-store promotional restrictions in the UK (DHSC, 2019), strategies to level the playing field are increasingly important. Strategies focusing on convenience, affordability and appeal are most likely to be successful among these groups (Food Foundation., 2020).

Outliers in the study reveal that the influence of deprivation may be moderated by education and ethnicity, while busy family lives could be an important barrier to purchasing fruit and vegetables. Outlier areas should be explored in more detail in subsequent studies to understand the local factors which cause them to buck the deprivation trend. This evidence would inform the current social prescribing debate by revealing local influencers of healthy diets. Further work should also explore whether diet-related inequalities are contributing to the spatial inequalities which can be observed in a wide range of health outcomes.

4.4.2 Future directions

Exploration of population diet using electronically captured secondary purchase data is in its relative infancy and, as such, we acknowledge several limitations which set out a foundation for future research. Future directions include; estimation of and controlling for household characteristics to extrapolate individual-level estimates; controlling for the inedible proportion of fruit and vegetables and food waste; estimating the fruit and vegetable content of composite dishes; exploring purchases of fruits and vegetables separately, breaking these down further by type; and exploring the effect of seasonality on purchasing behaviours. The validity of applying geographically weighted regression to neighbourhood level geographies, and the ability of existing survey data to completement supermarket purchase records for the development of small area estimation models, should also be considered.

4.5 Conclusions

In conclusion, supermarket loyalty card transactions allow us to investigate small area patterns in food purchase behaviours and reveal that areas purchasing fewer fruit and vegetable portions typically had younger residents, were less affluent, were closer to the supermarket but shopped less frequently, and had a lower total monthly spend with the retailer. In addition, we were able to unpack outliers such as those populated by students which had higher than expected fruit and vegetable purchases despite relative deprivation, illustrating that more nuanced relationships exist than those reported in earlier research.

4.6 Supplementary materials

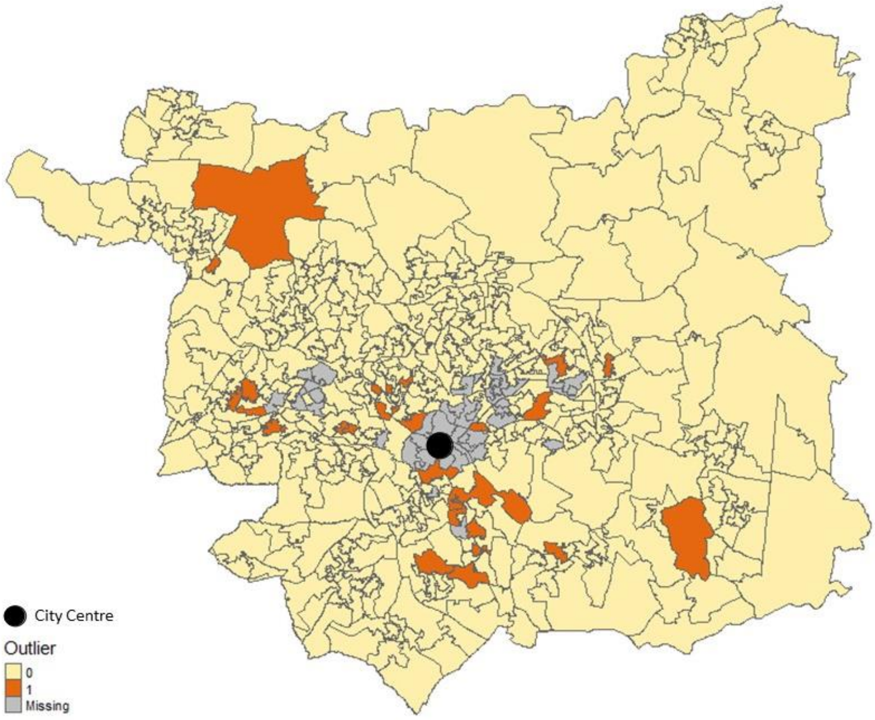
Table S1 - Outlier LSOAs (n=25) by IMD decile

IMD decile	Mean daily FV portions	% outlier customers
1	2.88	40.00%
2	2.82	8.00%
3	2.69	8.00%
4	2.97	12.00%
5	2.75	8.00%

6	3.27	8.00%
8	3.49	8.00%
9	2.77	8.00%
10	N/A	0.00%
TOTAL	2.93	100.00%

Figure S1 - Map of Outlier LSOAs from Model 1 (n=25)

1 indicates outlier areas according to Cooks Distance threshold 0.009



4.7 Additional information

Author Contributions: Conceptualization, V.J., M.A.M., G.P.C. and D.C.G. methodology, V.J., G.P.C., M.A.M. and D.C.G..; formal analysis, V.J.; resources, B.S., B.T. and T.R..; data curation, T.R., B.T., V.J. and M.A.M..; writing—original draft preparation, V.J.; writing—review and editing, V.J., M.A.M., G.P.C., D.C.G., T.R., B.T. and B.S.; supervision, M.A.M., D.C.G. and G.P.C.; funding acquisition, M.A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by an Economic and Social Research Council (ESRC) Data Analytics and Society Centre for Doctoral Training Grant, grant number ES/R501062/1.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the University of Leeds Ethics committee reference: AREA 18-050.

Informed Consent Statement: Informed consent was not required for this secondary data analysis, and not possible to obtain as all data were anonymized.

Data Availability Statement: Due to the commercial nature of the data used in this research, it is not possible for data to be published alongside the manuscript.

Acknowledgments: Thank you to the retailer for providing data in kind for this research; to Stephen Clark for technical expertise in GIS and R; to members of the lead author’s (VJ’s) Research Support Group panel for comments during the early stages of the work; and to Data Analytics team at the Leeds Institute for Data Analytics for support with data import to the secure environment.

Conflicts of Interest: T.R., B.T., and B.S. are employees at the grocery retailer. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

Adams, J., Goffe, L., Brown, T., Lake, A.A., Summerbell, C., White, M., Wrieden, W. and Adamson, A.J. 2015. Frequency and socio-demographic correlates of eating meals out and take-away meals at home: cross-sectional analysis of the UK national diet and nutrition survey, waves 1–4 (2008–12). *International Journal of Behavioral Nutrition and Physical Activity.* **12**(1), p51.

Aggarwal, A., Cook, A.J., Jiao, J., Seguin, R.A., Vernez Moudon, A., Hurvitz, P.M. and Drewnowski, A. 2014. Access to supermarkets and fruit and vegetable consumption. *American journal of public health.* **104**(5), pp.917-923.

Aiello, L.M., Quercia, D., Schifanella, R. and Del Prete, L. 2020. Tesco Grocery 1.0, a large-scale dataset of grocery purchases in London. *Scientific Data.* **7**(1), p57.

Appelhans, B.M., French, S.A., Tangney, C.C., Powell, L.M. and Wang, Y. 2017. To what extent do food purchases reflect shoppers' diet quality and nutrient intake? *International Journal of Behavioral Nutrition and Physical Activity.* **14**(1), p46.

Ball, K., Lamb, K.E., Costa, C., Cutumisu, N., Ellaway, A., Kamphuis, C.B.M., Mentz, G., Pearce, J., Santana, P., Santos, R., Schulz, A.J., Spence, J.C., Thornton, L.E., van Lenthe, F.J. and Zenk, S.N. 2015. Neighbourhood socioeconomic disadvantage and fruit and vegetable consumption: a seven countries comparison. *International Journal of Behavioral Nutrition and Physical Activity.* **12**(1), p68.

Birkin, M., Morris, M., Birkin, T. and Lovelace, R. 2018. Using census data in microsimulation modelling In: (eds), J.S.a.O.D.-W. ed. *The Routledge Handbook of Census Resources, Methods and Applications.* Routledge.

Blake, M.K. 2019. More than Just Food: Food Insecurity and Resilient Place Making through Community Self-Organising. *Sustainability.* **11**(10), p2942.

Bonevski, B., Randell, M., Paul, C., Chapman, K., Twyman, L., Bryant, J., Brozek, I. and Hughes, C. 2014. Reaching the hard-to-reach: a systematic review of strategies for improving health and medical research with socially disadvantaged groups. *BMC Medical Research Methodology.* **14**(1), p42.

Clark, S.D., Shute, B., Jenneson, V., Rains, T., Birkin, M. and Morris, M.A. 2021. Dietary Patterns Derived from UK Supermarket Transaction Data with Nutrient and Socioeconomic Profiles. *Nutrients.* **13**(5), p1481.

Clary, C., Lewis, D.J., Flint, E., Smith, N.R., Kestens, Y. and Cummins, S. 2016. The Local Food Environment and Fruit and Vegetable Intake: A Geographically Weighted Regression Approach in the ORiEL Study. *American journal of epidemiology.* **184**(11), pp.837-846.

DHSC. 2019. *Consultation on restricting promotions of products high in fat, sugar and salt*. Department of Health and Social Care. London: Assets Publishing Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/770704/consultation-on-restricting-price-promotions-of-HFSS-products.pdf

DHSC. 2020. *Tackling obesity: empowering adults and children to live healthier lives.* UK: Crown Copyright.

Drewnowski, A., Aggarwal, A., Cook, A., Stewart, O. and Moudon, A.V. 2016. Geographic disparities in Healthy Eating Index scores (HEI-2005 and 2010) by residential property values: Findings from Seattle Obesity Study (SOS). *Prev Med.* **83**, pp.46-55.

Eyles, H., Jiang, Y. and Ni Mhurchu, C. 2010. Use of Household Supermarket Sales Data to Estimate Nutrient Intakes: A Comparison with Repeat 24-Hour Dietary Recalls. *Journal of the American Dietetic Association.* **110**(1), pp.106-110.

Eyles, H., Jiang, Y. and Ni Mhurchu, C. 2010. Use of household supermarket sales data to estimate nutrient intakes: a comparison with repeat 24-hour dietary recalls. *Journal of the American Dietetic Association.* **110**(1), pp.106-110.

Food Foundation. 2020a. *Peas Please Progress Report 2020.* UK.

Food Foundation. 2020b. *Peas Please. Reviewing the evidence: what can retailers do to increase sales of fruit and veg.* London, UK.

Food Foundation. 2020c. *Veg Facts 2020: In Brief.* UK.

Food Foundation. 2021. *Veg Pledges.* [Online]. [Accessed 14/09/2021]. Available from: https://foodfoundation.org.uk/what-is-a-veg-city/veg-pledges/

Foresight. 2007. *Obesity Systems Map.* [Online]. [Accessed 13.06.2018]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/296290/obesity-map-full-hi-res.pdf

Fraser, L.K., Clarke, G.P., Cade, J.E. and Edwards, K.L. 2012. Fast Food and Obesity: A Spatial Analysis in a Large United Kingdom Population of Children Aged 13–15. *American Journal of Preventive Medicine.* **42**(5), pp.e77-e85.

Fraser, L.K., Edwards, K.L., Cade, J. and Clarke, G.P. 2010. The Geography of Fast Food Outlets: A Review. *International Journal of Environmental Research and Public Health.* **7**(5), pp.2290-2308.

Gale, C., Singleton, A., Bates, A. and Longley, P. 2016. Creating the 2011 area classification for output areas (2011 OAC). *Journal of Spatial Information Science.* **12**.

Gibson, S. and Neate, D. 2007. Sugar intake, soft drink consumption and body weight among British children: further analysis of National Diet and Nutrition Survey data with adjustment for under-reporting and physical activity. *Int J Food Sci Nutr.* **58**(6), pp.445-460.

GOV.UK. 2015. *English indices of deprivation 2015.* [Online]. [Accessed 11.09.2019]. Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015

GOV.UK. 2016. *National Diet and Nutrition Survey.* [Online]. Available from: https://www.gov.uk/government/collections/national-diet-and-nutrition-survey

Green, M.A., Watson, A.W., Brunstrom, J.M., Corfe, B.M., Johnstone, A.M., Williams, E.A. and Stevenson, E. 2020. Comparing supermarket loyalty card data with traditional diet survey data for understanding how protein is purchased and consumed in older adults for the UK, 2014–16. *Nutrition Journal.* **19**(1), p83.

Hamilton, S., Ni Mhurchu, C. and Priest, P. 2007. Food and nutrient availability in New Zealand: An analysis of supermarket sales data. *Public Health Nutrition.* **10**(12), pp.1448-1455.

Hauser, M., Nussbeck, F.W. and Jonas, K. 2013. The impact of food-related values on food purchase behavior and the mediating role of attitudes: A Swiss study. *Psychology & Marketing.* **30**(9), pp.765-778.

HMRC. 2018. *Check if your drink is liable for the Soft Drink Industry Levy.* [Online]. [Accessed 13.08.19]. Available from: https://www.gov.uk/guidance/check-if-your-drink-is-liable-for-the-soft-drinks-industry-levy

Jenneson, V., Pontin, F., Greenwood, D.C., Clarke, G.P. and Morris, M.A. 2021. A systematic review of supermarket automated electronic sales data for population dietary surveillance. *Nutrition Reviews.*

Leeds Observatory. 2019. *Full results for Leeds spreadsheet; IoD-2019-LSOA-Ward-Alt.xlsx.* [Online]. [Accessed 21.09.2021]. Available from: https://observatory.leeds.gov.uk/deprivation/

Leeds Observatory. 2021. *Household Size and Rooms in Leeds.* [Online]. [Accessed 28/09/2021]. Available from: https://observatory.leeds.gov.uk/housing/report/view/b535d8e5497342ebb3a4299f7284a6b7/E08000035/

Liese, A.D., Bell, B.A., Barnes, T.L., Colabianchi, N., Hibbert, J.D., Blake, C.E. and Freedman, D.A. 2014. Environmental influences on fruit and vegetable intake: results from a path analytic model. *Public health nutrition.* **17**(11), pp.2595-2604.

Maguire, E.R. and Monsivais, P. 2015. Socio-economic dietary inequalities in UK adults: an updated picture of key food groups and nutrients from national surveillance data. *Br J Nutr.* **113**(1), pp.181-189.

McCartney, G. 2011. Illustrating health inequalities in Glasgow. *Journal of Epidemiology and Community Health.* **65**(1), pp.94-94.

Menezes, M.C., Costa, B.V., Oliveira, C.D. and Lopes, A.C. 2017. Local food environment and fruit and vegetable consumption: An ecological study. *Prev Med Rep.* **5**, pp.13-20.

Mills, S., Adams, J., Wrieden, W., White, M., Brown, H.,. 2018. Sociodemographic characteristics and frequency of consuming home-cooked meals and meals from out-of-home sources: cross-sectional analysis of a population-based cohort study. *Public Health Nutrition.*

Morris, M., Clarke, G., Edwards, K., Hulme, C. and Cade, J. 2016. Geography of Diet in the UK Women’s Cohort Study: A Cross-Sectional Analysis. *Epidemiology - Open Journal.* **1**(1), pp.20-32.

Morris, M.A., Hulme, C., Clarke, G.P., Edwards, K.L. and Cade, J.E. 2014. What is the cost of a healthy diet? Using diet data from the UK Women's

Cohort Study. *Journal of Epidemiology and Community Health.* **68**(11), pp.1043-1049.

Närhinen, M., Berg, M.-A., Nissinen, A. and Puska, P. 1999. Supermarket sales data: a tool for measuring regional differences in dietary habits. *Public Health Nutrition.* **2**(3), pp.277-282.

National Food Strategy. 2020. *National Food Strategy: Part One.* UK.

NDNS. 2018. National Diet and Nutrition Survey (NDNS), years 7 and 8 (2014/15 - 2015/16).

Nestle, M. and Jacobson, M.F. 2000. Halting the obesity epidemic: a public health policy approach. *Public Health Reports.* **115**(1), pp.12-24.

NHS. 2018a. *5 A Day: what counts?* [Online]. [Accessed 22/06/2020]. Available from: https://www.nhs.uk/live-well/eat-well/5-a-day-what-counts/

NHS. 2018b. *Why 5 a day?* [Online]. [Accessed 10.09.2019]. Available from: https://www.nhs.uk/live-well/eat-well/why-5-a-day/

NHS. 2019. *Obesity.* [Online]. [Accessed 10.09.2021]. Available from: https://www.nhs.uk/conditions/obesity/

NHS Digital. 2019. *Health Survey for England 2018 overweight and obesity in adults and children.* London: Health and Social Care Information Centre.

Office for National Statistics. 2013. *2011 Census: Population estimates by single year of age and sex for local authorities in the United Kingdom: Unrounded estimates of the usually resident population by age and sex, along with household estimates on census day, 27 March 2011.* [Online]. [Accessed 21.09.2021]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/2011censuspopulationestimatesbysingleyearofageandsexforlocalauthoritiesintheunitedkingdom

Office for National Statistics. 2021. *Census Output Area population estimates- Yorkshire and The Humber, England (supporting information).* [Online]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/censusoutputareaestimatesintheyorkshireandthehumberregionofengland

Ogden, J., Bandara, I., Cohen, H., Farmer, D., Hardie, J., Minas, H., Moore, J., Qureshi, S., Walter, F. and Whitehead, M.-A. 2001. General practitioners and patients models of obesity: whose problem is it? *Patient Education and Counseling.* **44**(3), pp.227-233.

ONS. 2017a. *Census geography: An overview of the various geographies used in the production of statistics collected via the UK census.* [Online]. [Accessed 23.12.17]. Available from: https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography

ONS. 2017b. *Living costs and food survey*. Bulman. Jo. User guidance and technical information for the Living Costs and Food Survey. UK: ONS. Available from:

https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/methodologies/livingcostsandfoodsurvey

Osbourne, B., Cooper, V. and Neave, A. 2018. *Health Survey for England 2017 Adult health related behaviours.* NHS Digital.

PHE. 2014. *Obesity and the environment: regulating the growth of fast food outlets.* London, England.

Public Health England. 2016. *National Diet and Nutrition Survey.* [Online]. [Accessed 20/05/2019].

Public Health England. 2019. *NDNS: time trend and income analyses for Years 1 to 9.* Public Health England., and the Food Standards Agency.,.

Rains, T. and Longley, P. 2021. The provenance of loyalty card data for urban and retail analytics. *Journal of Retailing and Consumer Services.* **63**, p102650.

Ransley, J.K., Donnelly, J.K., Khara, T.N., Botham, H., Arnot, H., Greenwood, D.C. and Cade, J.E. 2001. The use of supermarket till receipts to determine the fat and energy intake in a UK population. *Public Health Nutrition.* **4**(6), pp.1279-1286.

Rehm, J., Kilian, C., Rovira, P., Shield, K.D. and Manthey, J. 2021. The elusiveness of representativeness in general population surveys for alcohol. *Drug and Alcohol Review.* **40**(2), pp.161-165.

Robards, J., Gale, C. and Martin, D. 2017. *Creating a synthetic spatial microdataset for zone design experiments using 2011 Census and linked administrative data.*

Scarborough, P., Allender, S., Peto, V., and Rayner, M. . 2008. *Regional and social differences in coronary heart disease.* Oxford, UK: British Heart Foundation Health Promotion Research Group. Department of Public Health, University of Oxford.

Theis, D.R.Z. and White, M. 2021. Is Obesity Policy in England Fit for Purpose? Analysis of Government Strategies and Policies, 1992–2020. *The Milbank Quarterly.* **99**(1), pp.126-170.

Vepsalainen, H., Nevalainen, J., Kinnunen, S., Itkonen, S.T., Meinila, J., Mannisto, S., Uusitalo, L., Fogelholm, M. and Erkkola, M. 2021. Do we eat what we buy? Relative validity of grocery purchase data as an indicator of food consumption in the LoCard study. *Br J Nutr.* pp.1-24.

Vickers, D., Rees, P. and Birkin, M. 2005. *Creating the National Classification of Census Output Areas: Data, Methods and Results* School of Geography Working Paper 05/: University of Leeds. Available from: http://eprints.whiterose.ac.uk/5003/

WCRF. 2018. *Diet, Nutrition, Physical Activity and Cancer: a Global Perspective.* World Cancer Research Fund/American Institute for Cancer Research.

Winkler, E. and Turrell, G. 2010. Confidence to cook vegetables and the buying habits of Australian households. *J Am Diet Assoc.* **110**(5 Suppl), pp.S52-61.

Wrap. 2018. *Household food waste: restated data for 2007-2015.*

Yau, A., Adams, J. and Monsivais, P. 2019. Time trends in adherence to UK dietary recommendations and associated sociodemographic inequalities, 1986-2012: a repeated cross-sectional analysis. *Eur J Clin Nutr.* **73**(7), pp.997-1005.

Chapter 5
Supermarket Transaction Records In Dietary Evaluation – The STRIDE study: validation against self-reported dietary intake

Victoria Jenneson, Darren C Greenwood, Graham P Clarke, Tim Rains, Bethan Tempest, Becky Shute, Michelle A Morris

In preparation for Public Health Nutrition

Keywords: dietary assessment, validation, macronutrients

Abstract

Objective: Scalable methods are required for population dietary monitoring. The STRIDE study compares dietary estimates from supermarket transactions with an online Food Frequency Questionnaire.

Design: Purchases were collected for one year during the study and one year prior, across four cohorts. Bland-Altman agreement and limits of agreement (LoA) were calculated for energy, sugar, fat, saturated fat, protein and sodium (absolute and energy-adjusted).

Setting: This study was in partnership with a large UK retailer.

Participants: 1,788 participants from four UK regions were recruited from the retailer's loyalty card customer database, according to their breadth and frequency of purchases. 686 participants were included for analysis.

Results: Participants were mostly female (72%), with a mean age of 56 years (SD 13). The ratio of purchases to intakes varied, depending on amounts purchased and consumed, with purchases under-estimating intakes for smaller amounts on average, but over-estimating for larger amounts.

For absolute measures, the limits of agreement across households were wide, e.g. for energy intake of 2000kcal, purchases could under or over-estimate intake by a factor of 5; values could be between 400kcal to 10000kcal. Limits of agreement for relative estimates were smaller, e.g. for 14% of total energy from saturated fat, purchase estimates may be between 7% and 27%.

Conclusions: Agreement between purchases and intake was strongest for smaller households, loyal customers, and for energy-adjusted values.

Purchases are a good proxy for dietary composition and useful for population dietary surveillance, ecological studies, and identifying intervention targets.

5.1 Introduction

National dietary surveys, such as the UK's National Diet and Nutrition Survey (NDNS) (Public Health England., 2019) can reveal dietary trends that impact health. However, costs and administrative burdens associated with surveys limit their sample sizes and temporal granularity, due to their cross-sectional nature. Online food records such as myfood24 (Carter et al., 2016) and intake24 (Simpson et al., 2017) improve scalability and reduce costs associated with dietary surveys (Burley et al., 2015). Yet, digital methods continue to rely on self-report which is known to exhibit social desirability and recall biases leading to under-estimation of energy intake (Ravelli and Schoeller, 2020). Harnessing new technology could benefit research by providing a suite of scalable objective dietary assessment methods to complement existing self-report (de la Hunty et al., 2021). Objective dietary measures may come in the form of image capture techniques (Buttriss et al., 2017), nutritional biomarkers (Buttriss et al., 2017), and food system administrative data such as transaction records (Jenneson et al., 2021).

Food purchases represent upstream dietary behaviours which precede consumption. Advancements in technology now permit the routine collection of purchase data by supermarkets in the form of Electronic Point of Sale (EPOS). EPOS data is used commercially for stock analysis, customer segmentation and marketing. In combination with product nutritional information and customer data, EPOS transactions could provide objective population-level dietary insight at a much larger scale (Nevalainen et al., 2018). As such, researchers have begun to explore the value of supermarket electronic purchase records in a dietary research context (Jenneson et al., 2021; Eyles et al., 2010; Green et al., 2020; Vepsalainen et al., 2021; Appelhans et al., 2017).

A precursor to digital purchase records was the collection of paper till receipts, often accompanied by purchase diaries. Early work in the UK found paper till receipts demonstrated good statistical agreement with self-reported individual consumption (Ransley et al., 2001). Yet scalability of the method was limited by the need for participants to collect their receipts and the burden of manual coding by researchers. While paper receipts enable purchases from different sources to be combined, there remains the potential for participants to lose or systematically omit receipts. With the ability to eliminate these burdens and biases, pioneering work by Ransley et al. (2001) demonstrates promise for digitised receipt collection methods in dietary assessment.

Automatically captured electronic supermarket transaction records are becoming increasingly employed in dietary research and monitoring of dietary policy (Jenneson et al., 2021), thanks to their scale, timeliness and richness of detail. However, with only a few studies to date investigating their validity as a dietary measure there is a need for further validation studies. One such study by (Eyles et al., 2010) included just 49 customers of a New Zealand supermarket, and 3-months of transaction records. Comparison of nutrients from household transaction records with self-reported intake from four random 24-hour dietary recalls, revealed differing strengths of correlation by nutrient. Similarly, in a study comparing grocery purchases with a food frequency questionnaire (FFQ) for nearly 12,000 Finnish loyalty card holders, strength of association at the food group level varied substantially (Vepsalainen et al., 2021).

These previous validation studies point to the favourable use of household supermarket transaction records to act as a proxy for individual dietary intake, but suggest that utility of the method is likely to depend upon the food group or nutrient in question. Furthermore, while household composition and retailer loyalty appear to be important factors (Vepsalainen et al., 2021), neither study attempted to account for household composition in their estimates, presenting an area for additional research. Comparisons with self-reported intake should also include alcohol, food waste, and consumption by visitors to improve agreement (Eyles et al., 2010). Self-reported diet needs to cover more days, and with larger sample sizes, to allow for large-intra-individual variation, particularly for sugar and total energy (Vepsalainen et al., 2021).

This paper presents results from the STRIDE study (Supermarket Transaction Records In Dietary Evaluation) (Jenneson et al., 2020), which adds novel insight to the existing evidence by assessing the statistical agreement between estimates of nutrient purchases from loyalty card transaction records and estimates of nutrient intake from an online FFQ. The present study adds to existing knowledge by assessing statistical agreement, and limits to agreement for both absolute and energy-adjusted macronutrient estimates, and accounting for household composition to derive individual-level estimates of purchased nutrients.

5.2 Methods

The study protocol (Jenneson et al., 2020) was registered prior to starting the study on the Open Science Framework and is available online at https://doi.org/10.17605/OSF.IO/VUKTQ. The protocol received approval from the University of Leeds ethical review board (AREA/18-174).

5.2.1 Study design

This validation study compares self-reported intake against household food purchase data from a major UK retailer's loyalty card scheme. Intake is captured for the previous 2-3 months using an online FFQ by the Scottish Collaborative Group (SCG) (Masson et al., 2003). The study recruited four study cohorts (plus a pilot phase) and was designed to capture intake and purchase data across all seasons. Transactions cover a 1-year baseline period prior to study recruitment and a 1-year period during which the STRIDE study took place. Transaction data which cover the same 3 months as that which is captured by the FFQ for each cohort, is referred to as the 'cohort period'. The period covered by each cohort is depicted in Figure 5.1. More detail on the participant recruitment is provided below.

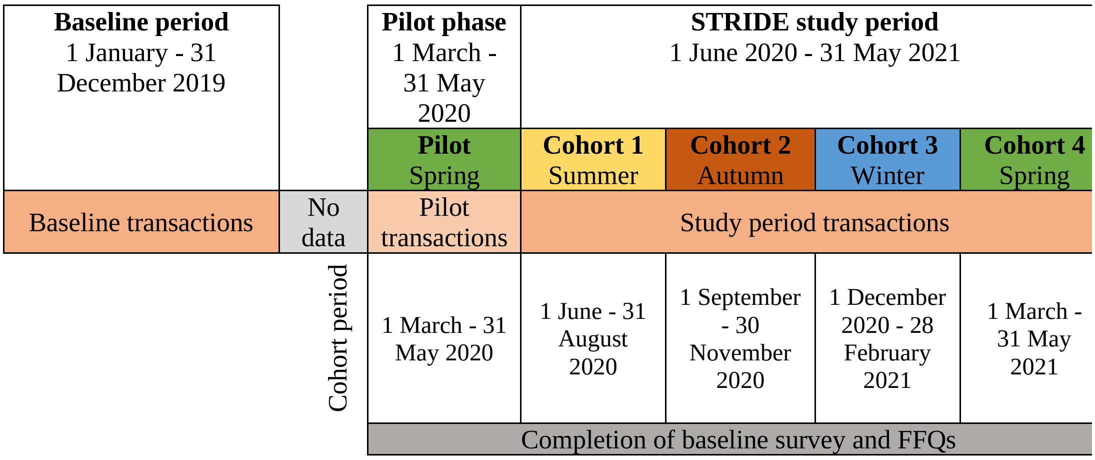


Figure 5.1 STRIDE study design

5.2.2 Participant sampling and recruitment

Participants were sampled from the retailer's database of loyalty card holders. Customers must be at least 18 years old to hold a loyalty card. Eligible customers were required to have an email address on file, to have opted in to receive research communications, and to have their loyalty card registered to an address in one of four regions in England (Yorkshire and the Humber, South-East, East Midlands, West Midlands), selected to cover a range of geographic and demographic characteristics. Primary shoppers were selected, for whom we considered purchases at the study retailer likely to represent the majority of their shopping. This was determined by selecting only those customers who shopped in at least seven out of 15 food categories on a minimum of 10 occasions during the 2019 calendar year. Additionally, customers with an annual spend on food and non-alcoholic beverages greater than 1.5 times that published in the 2019 edition of the Family Food Survey (FFS) (GOV.UK, 2020) were excluded. the same proportion was also excluded from the lower end of the distribution of annual spend.

A pilot phase was used to determine the expected sign-up and completion rates for the study and to test the flow of the participant journey. This indicated a sign-up rate of around 1% and a completion rate for the FFQ of around 50%. To achieve a sample size of 200 customers per cohort for statistical agreement testing, all customers who met the eligibility criteria (~45,000) were invited by the retailer via email to take part in one of the STRIDE study cohorts. The participant journey is shown in Figure 5.2. Customers received an invitation email containing two links, one to the online consent form and baseline questionnaire hosted by Jisc Online Surveys, and the other to the study website (hosted by the Leeds Institute for Data Analytics (LIDA)) where participant information could be found.

Customers consented to receive a further email containing a link to the SCG online FFQ, and to allow their loyalty card purchase records for 1-year prior to, and 1-year during the study to be shared with the research team. Customers provided their loyalty card number to enable their purchase records to be identified by the retailer. A unique customer identifier was also embedded into the URL in the invitation email to aid identification in the event of typos in loyalty ID data entry by participants. Purchase records were linked to the FFQ and each participant’s baseline survey and dietary questionnaire via a unique study ID assigned to each participant. Upon completion of the FFQ, customers were entered into a prize draw for a chance to win a £75 high street voucher (one per cohort including the pilot phase) as an incentive to participate in the study.

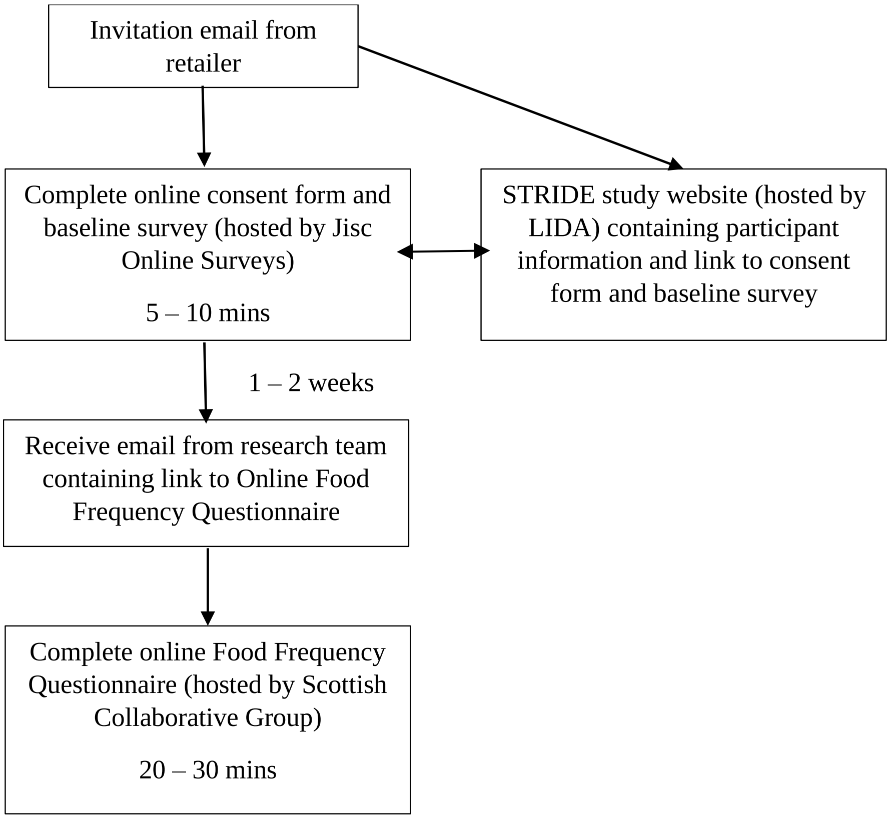

Invitation email from retailer
Complete online consent form and baseline survey (hosted by Jisc Online Surveys)
5 – 10 mins
STRIDE study website (hosted by LIDA) containing participant information and link to consent form and baseline survey
1 – 2 weeks
Receive email from research team containing link to Online Food Frequency Questionnaire
Complete online Food Frequency Questionnaire (hosted by Scottish Collaborative Group)
20 – 30 mins

Figure 5.2 STRIDE participant journey flow diagram

5.2.3 Data collection

Demographic information (date of birth, gender, ethnicity, and height and weight for calculation of body mass index (BMI)) were collected via an online baseline questionnaire. Participants additionally reported: the number and ages of other people in their household; the proportion of their food purchases made with the retailer by selecting one of five categories on the baseline questionnaire (0-20%, 20-40%, 40-60%, 60-80%, or 80-100%); dietary restrictions; food waste; and the impacts of the COVID-19 pandemic on their food purchase and consumption habits.

Food consumption data was collected via the SCG Online FFQ (Masson et al., 2003), a 150-item semi-quantitative questionnaire which asks the participant to report the frequency and amounts consumed for each item. Transaction data and product nutrition information were provided by the retailer.

All food and beverages (including alcoholic beverages) purchased either in store or online with a scanned loyalty card were recorded. Transaction files contained a row for each product (with a unique product ID) with an item description, purchase quantity (units or weight as appropriate) and cost (GBP

£). Products purchased on a single shopping trip may be linked by a transaction ID, thus a transaction represents a basket of goods.

5.2.4 Nutrient estimates

Daily nutrient intakes for each participant were estimated from their FFQ by the SCG team as part of their paid FFQ service, using the UK National Nutrient Databank (Masson et al., 2003). Purchased nutrients were estimated from the transaction data by linking products to a bespoke product nutrient composition database (PNCD) via a unique product code (either the European Article Number (EAN) or Stock-keeping Unit (SKU)). The PNCD comprised of back of pack product nutrient information per 100g or per 100ml of product ((Energy (kcal), total sugars (g), protein (g), total fat (g), saturated fat (g), and sodium (mg)) provided by the retailer for products sold in 2019. 72% of products were matched to product-specific nutrient data in the retailer file. For products where no match could be found in the nutrient file, or where nutrition information was blank, generic values were imputed from the UK's Composition Of Food Integrated Dataset (CoFID) version 7 (PHE, 2020). This was typically for non-packaged items such as fresh produce and in-store bakery items, alcohol (for which nutritional information is not legally required to be displayed on product packaging (Department of Health., 2017)), and seasonal products such as Easter eggs. After imputation a match rate of 100% was achieved.

The product weight was multiplied by the number of units purchased and its nutritional value per 100g (or 100ml; as specific gravity information was unavailable for products a simple approximation of 1ml = 1g was assumed) to derive the total nutrients purchased in a given period by each customer. For comparison with daily intake estimates, purchased nutrients were converted to mean daily household estimates by dividing the total nutrients purchased by the number of days in the cohort period (covering the same 3-month timeframe as each FFQ). Individual-level daily purchase estimates were generated from household estimates by allocating purchased nutrients to individuals proportionate to UK dietary recommendations for energy intake by age and gender (PHE, 2016) (Table 5.1). As genders were unknown for other household members, an average of recommended values for females and males was used. For example, if a study participant is a 30-year-old woman living with a 30-year-old partner and a 3-year-old child, she would be allocated 36% of the nutrients purchased by the household (1928/(1928 + 2230 + 1197.5)) (Table 5.1).

Table 5.1 UK recommended daily calorie intakes by age and gender Source; PHE, 2016

	Recommended daily energy intake (kcal)	
Age (years)	**Female**	**Male**
0 – 1	698	745
1 – 3	1165	1230
4 – 10	1656	1861
11 – 17	1959	2449
18 – 64	1928	2532
65+	1855	2215

Estimates are given both as absolute daily amounts and energy-adjusted values. Macronutrients are expressed in terms of their contribution to total energy by multiplying the number of calories per gram (protein = 4 kcal/g, fat = 9 kcal/g, saturated fat = 9 kcal/g, sugars = 3.9 kcal/g) by the number of grams, then dividing by total energy. Sodium is expressed as mg/kcal. Absolute estimates from purchase data are given at the household-level and individual-level, whilst energy-adjusted purchase estimates are presented as a single figure. This is because the same proportions are used to allocate energy and all other nutrients to household members, thus the individual-level estimate would be the same as the household-level estimate.

5.2.5 Statistical analysis

Daily nutrient purchase estimates (at the household-level and at the individual-level) for each cohort period are compared with individual-level nutrient intakes for the same time period. Purchased energy at the individual-level is compared with energy intake (split by household size and self-reported customer loyalty). Due to low numbers, customers with a household size of three or more were combined, and compared with single-person and two-person households. Similarly, customers reporting that the retailer contributes 0-20%, 20-40%, or 40-60% of their food purchases were combined to represent low-medium loyalty customers (0-60% of food purchases), and compared with high loyalty customers (60-80% of food purchases) and very high loyalty customers (80-100% of food purchases). Individual-level purchase estimates for macronutrients and sodium are compared with intake estimates.

Bland-Altman plots were generated to assess statistical agreement and limits of agreement (LoA) (Bland and Altman,1986; Bland and Altman, 1999). Due to heteroskedasticity in the data (the difference between measures was related to the magnitude of the mean of the measures), values in the Bland-Altman plots are log-transformed. The axes of the Bland-Altman plots are back-transformed to aid interpretation and shown as a ratio of purchase estimate/intake estimate against the mean (Bland and Altman, 1999). This ratio may then be interpreted as a percentage difference. As the direction of the relationship was also dependent upon the magnitude of measures, a regression approach was used to plot the mean difference (which is presented as a regression equation in the tables) and limits of agreement, based on ±1.96 standard deviations of the spread of residuals about the regression line (Bland and Altman, 1999; Bland, 2005).

5.3 Results

5.3.1 Participant characteristics

Recruitment figures for the STRIDE study are shown in Table 5.2. Around half of the 1,788 participants recruited across the whole study completed an online FFQ (n=825). Of those with completed FFQ records, 83% (n=688) had made at least one purchase with the retailer in the corresponding 3-month period as covered by the FFQ. A further two participants were excluded as outliers; their estimated daily energy intake from the FFQ was ≥8,000kcal (four times the recommended calorie intake for an adult woman). Results presented are pooled across all cohorts (including the pilot) for those 686 participants.

Table 5.2 STRIDE participant recruitment summary

Cohort	**Pilot**	**Cohort 1**	**Cohort 2**	**Cohort 3**	**Cohort 4**	**All cohorts combined**
Number of participants consented	80	377	547	430	354	1788
Number of participants with	38 (48)	190 (50)	235 (43)	192 (44)	170 (48)	825 (46)

completed FFQs (%)						
Analysis sample. (Number of participants with completed FFQs and purchase data in the cohort period (%))	13 (16)	159 (42)	201 (37)	159 (37)	156 (44)	688 (38)

The demographic characteristics of study participants are shown in Table 5.3. The majority of participants were female (72%) and from a white ethnic background (97%). Participants had a mean age of 56.2 years (standard deviation 12.9 years), and an average household size of 2.2 persons. According to their self-reported height and weight, 54% of participants were classified as overweight or obese. 30% had a loyalty card registered to an address in the Yorkshire and Humber region, 20% in the East Midlands, 18% in the West Midlands, and 30% in the South East. Participants were relatively affluent overall with almost 69% living in areas in the five least deprived deciles according to the Index of Multiple Deprivation (IMD) (GOV.UK, 2015), and the most commonly inhabited Output Area Classification Supergroup areas (Gale et al., 2016) were Suburbanites (31%), Urbanites (27%) and Rural Residents (17%). Participants were also relatively loyal to the study retailer, with 82% reporting to purchase at least 40% of their food and beverages with the retailer and 64% purchasing at least 60%.

Overall, there was little difference in the characteristics of those who signed up for the study and those included in the analysis sample. Small observed differences include a smaller proportion of non-white participants, a higher proportion of healthy weight and a lower proportion of obese individuals, a smaller proportion of Constrained City Dwellers and a higher proportion of Rural Residents, as well as a slight reduction in household size in the analysis sample compared with total sign-ups.

Table 5.3 STRIDE participant characteristics for all cohorts (including pilot) combined

Participant characteristics	**Number of consented participants (n = 1788)**	**Number of participants with completed FFQs (n = 825)**	**Analysis sample** **(Number of participants with completed FFQs and purchases recorded in the cohort period) (n = 686)**
Gender (%)			
Female	1,303 (73)	597 (72)	497 (72)
Male	479 (27)	224 (27)	186 (27)
Other/unknown	6 (0)	4 (1)	3 (0)
Age, mean (SD)	56 (14)	57 (13)	56 (13)
Ethnicity (%)			
White	1,711 (96)	803 (97)	667 (97)
Non-white	52 (3)	12 (2)	9 (1)
Other/unknown	25 (1)	10 (1)	10 (1)
BMI (kg/m^2)			
Underweight	187 (11)	91 (11)	73 (11)
Healthy	496 (28)	249 (30)	215 (31)
Overweight	513 (29)	241 (29)	198 (29)
Obese	382 (21)	163 (20)	135 (20)
Morbidly obese	112 (6)	50 (6)	38 (6)
Government Office Region			
Yorkshire and The Humber	519 (29)	249 (30)	209 (31)
East Midlands	338 (19)	161 (20)	135 (20)
West Midlands	360 (20)	146 (18)	126 (18)
South East	484 (27)	233 (28)	206 (30)

Other/unknown	87 (5)	36 (4)	10 (2)
IMD Decile			
1 – most deprived	73 (4)	31 (4)	27 (4)
2	96 (5)	31 (4)	25 (4)
3	114 (6)	46 (6)	38 (6)
4	136 (8)	59 (7)	49 (7)
5	183 (10)	85 (10)	74 (11)
6	188 (11)	106 (13)	85 (12)
7	202 (11)	91 (11)	76 (11)
8	221 (12)	103 (13)	95 (14)
9	223 (13)	96 (12)	85 (12)
10 – least deprived	281 (16)	150 (18)	130 (19)
Unknown	71 (4)	27 (3)	2 (0)
OAC Supergroup (2011)			
-Rural Residents	254 (14)	142 (17)	116 (17)
-Cosmopolitans	45 (3)	27 (3)	23 (3)
-Ethnicity Central	15 (1)	5 (1)	4 (1)
-Multicultural Metropolitans	124 (7)	55 (7)	46 (7)
-Urbanites	465 (26)	208 (25)	183 (27)
-Suburbanites	522 (29)	249 (30)	216 (32)
-Constrained City Dwellers	75 (4)	20 (2)	16 (2)
-Hard-pressed Living	217 (12)	92 (11)	80 (12)
Unknown	80 (5)	27 (3)	2 (0)
Share of purchases made			

with study retailer (%)			
0 – 20%	122 (7)	58 (7)	44 (6)
21 – 40%	239 (13)	97 (12)	81 (12)
41 – 60%	309 (17)	145 (18)	119 (17)
61 – 80%	456 (26)	195 (24)	174 (25)
81 – 100%	661 (37)	330 (40)	268 (39)
Mean household size (SD)	2.3 (1.3)	2.2 (1.1)	2.2 (1.1)

5.3.2 Descriptive statistics

Absolute daily estimates of purchased (household-level and individual-level) and consumed energy and nutrients are presented in Table 5.4. Household purchase estimates are around 80-90% of the consumed estimate value, depending on the nutrient. Individual purchase estimates are around half the amount purchased at the household level. Individual-level purchase estimates are around 40% of the estimated consumption amount.

Table 5.4 Absolute nutrient estimates from purchase records and food frequency questionnaires (n = 686)

Nutrient	**Absolute household purchase/day**		**Absolute individual-level purchase/day**		**Absolute consumption/day (FFQ)**	
	Median	IQR	Median	IQR	Median	IQR
Energy (kcal)	1746	803, 3233	910	371, 1621	1955	1584, 2480
Sugar (g)	82	35, 162	42	17, 83	107	83, 145
Protein (g)	65	27, 117	33	13, 60	83	65, 104
Total fat (g)	72	31,133	37	15, 66	79	61, 102
Saturated fat (g)	27	12, 52	14	6, 26	31	23, 41
Sodium (mg)	1984	781, 3661	1031	403, 1892	2623	2090, 3374

Energy-adjusted nutrient estimates are given for purchases and consumption in Table 5.5. Participants purchased on average 19% of their energy from sugar, 14% from protein, 36% from fat, 14% from saturated fat, and an average of 1.06 mg sodium per calorie.

Table 5.5 Energy-adjusted nutrient estimates from purchase records and food frequency questionnaires (n = 686)

Nutrient	**Energy-adjusted purchase/day**		**Energy-adjusted consumption/day (FFQ)**	
	Median	IQR	Median	IQR
Sugar (% energy)	19	16, 23	21	18, 25
Protein (% energy)	14	12, 16	17	15, 19
Total fat (% energy)	36	32, 41	37	33, 40
Saturated fat (% energy)	14	12, 16	14	12, 16
Sodium (mg/kcal)	1.1	0.9, 1.3	1.3	1.2, 1.5

5.3.3 Agreement for absolute estimates

The relationship between absolute daily purchases and absolute daily intake was examined for calories. As shown in the scatterplot in Figure 5.3a), correlation between the two measures is weak (Pearson’s correlation coefficient = 0.02). Correlation tells us to what extent measures follow the same linear pattern, this is not the same as agreement which tells us the average magnitude of difference between measures. Agreement between daily household energy purchased and daily energy intake by the participant can be seen in the Bland-Altman plot in Figure 5.3b), which plots the mean of the measures against the difference between them. The horizontal black line (line of equality) indicates perfect agreement (difference = 0) between measures. The blue horizontal line shows the arithmetic mean difference across all data points. The dashed lines show the 95% limits of agreement (LoA) around the mean difference. The datapoints do not cluster neatly around the line of equality, demonstrating evidence of heteroskedasticity, that is, the difference between measures varies with the magnitude of the mean, in both directions.

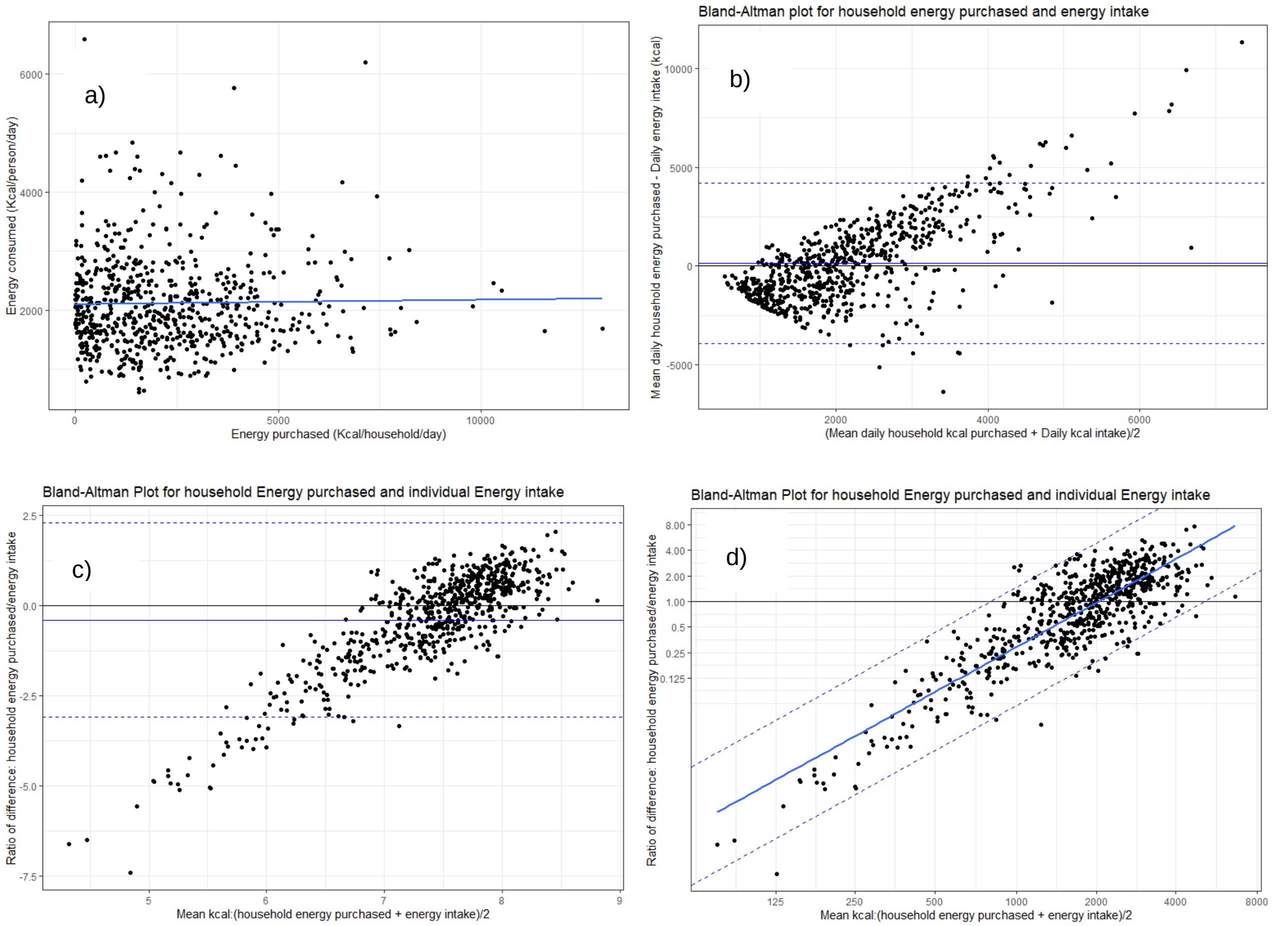


Figure 5.3 Household energy purchased vs daily energy intake (kcal) a) scatterplot, b) Bland-Altman plot for agreement, c) Bland-Altman plot for log-transformed variables, d) Bland-Altman plot for log-transformed variables with regression approach, with difference expressed as a ratio of purchases:intake.

As advised by Bland and Altman (1999), the data were log-transformed to account for heteroskedasticity (Figure 5.3c)). Here the mean of log household energy purchased and log individual energy intake is plotted against the difference between log household energy purchased and log individual energy intake. At lower magnitudes of energy, purchased household energy is lower than energy intake, while at higher magnitudes, purchased household energy is higher than energy intake. Thus, the arithmetic mean difference and derived LoAs (Supplementary Table 1) do not represent well the agreement across the distribution and should be interpreted with caution.

The agreement and LoAs are more appropriately shown as regression lines (Figure 5.3d)). The β_o and β_1 coefficients which make up the regression equations can be found in Table 5.6. The Bland-Altman plot in Figure 5.3d), shows the agreement for log-transformed variables, with the axes labels back-transformed to aid interpretation. Thus, the x-axis can be interpreted as the mean between household energy purchase and individual energy intake (in kcals) and the y-axis as the difference as a ratio of purchased energy to energy intake. The line of equality (horizontal black line) is now represented by 1 (a 1:1 ratio between measures representing 100% agreement). Values greater than 1 indicate that purchase estimates are higher than intake, while values lower than 1 indicate purchase estimates are lower than intake.

Figure 5.3d) shows that average agreement between household energy purchased and individual energy intake is near perfect where the mean of values is around 2000kcal. Yet, the shape of the line indicates that below this magnitude purchased energy is likely to be lower than energy intake, while above this magnitude purchased energy is likely to be higher than energy intake. Taking the intercept and slope of the regression lines (Table 5.6), it is therefore possible to estimate the expected agreement for a given magnitude. For example, for an average daily intake of 2000kcal (A), the natural log of A (7.6) is multiplied by the slope, then added to the intercept to give the log of the difference, which is back-transformed to give the ratio of purchase:intake. Results suggest that at a mean of 2,000kcal, household purchases under-estimate individual calorie intake by just 2% (~40kcal) on average. Yet the wide limits of agreement mean that household energy purchased could be anywhere from just 20% of calorie intake, to almost 5 times higher, demonstrating a lack of confidence in the agreement estimate.

It was observed that a number of customers in our sample had very low daily calorie purchases, which may be influencing our agreement results. Therefore a sensitivity analysis was conducted excluding customers who purchased less than 500kcal/day on average (n=124). A cut-off of 500kcal was chosen as this represents a quarter of the daily recommended calorie intake for an adult woman, mirroring our upper cut-off of 8000kcal, which represents four-times the recommended intake. Results for the sensitivity analysis (Table 5.6) show that exclusion of the lowest purchasing customers does not change the mean difference much at an intake of 2000kcal, yet this time household purchases over-estimate intake by around 2% on average. Additionally, the slope of the line is reduced and limits of agreement are narrower (as seen on the charts in Supplementary Figure 1), indicating closer agreement is observed where the least loyal customers are excluded.

Accounting for household composition, individual-level purchased energy (Supplementary Figure 2) under-estimates intake by around 14% (A = 2000kcal), yet limits of agreement are narrower than for household purchases (agreement = 86%, LoA 22% – 343%). Exploration of sub-groups by household size (Supplementary Figure 3) show that (for A = 2000kcal) purchased energy estimates are closest to intake estimates for single-person (agreement = 98%, LoA 23% - 393%) and two-person households (agreement = 99%, LoA 27% - 365%), but further for larger households containing three or more persons (agreement = 91%, LoA 21% - 387%). For single-person households, household-level and individual-level estimates are equivalent). Yet limits of agreement remain wide, suggesting that purchases are likely to under and over-estimate intake. There is also an association between agreement of measurements and customer loyalty (Supplementary Figure 4). For customers reporting a low-medium loyalty with the retailer (0-60% of their food shopping), for an average intake of 2000kcal individual-level energy purchase estimates tend to under-estimate intake, representing 82% of intake on average (LoA 21%, 325%), while in the most loyal customer group (80-100% of food shopping carried out with the retailer) individual-level purchase estimates over-estimate calorie intake (agreement = 113%, LoA 30%, 431%).

Absolute daily purchases at the individual level were also compared to intake for macronutrients and sodium (results not presented). To summarise, the nutrients showed similar patterns to those observed for energy; variance in agreement with magnitude of the mean of measures; a tendency for purchases to over-estimate intake at the top end of the distribution and to under-estimate intake at the lower end; and wide limits of agreement. Thus, our results suggest that for all examined nutrients, purchase data provides a poor proxy of individual intake, even when adjusted for household composition.

Table 5.6 Regression coefficients for mean difference and limits of agreement between purchase and intake for energy (kcal)

	Mean difference, (purchase / intake)			**Lower limit of agreement**		**Upper limit of agreement**	
	Intercept *b₀*)	**Slope (*b1*)**	**Ratio of difference A = 2000**	**Intercept *b₀*)**	**Ratio of difference A = 2000**	**Intercept *b₀*)**	**Ratio of difference A = 2000**
Household purchase – Intake							
All households (n = 686)	-13.25	1.74	0.98	-14.86	0.20	-11.64	4.88
Sensitivity analysis (n = 562)	-8.19	1.08	1.02	-9.46	0.29	-6.93	3.61
Individual purchase – Intake (by household size)							
All households (n = 686)	-13.60	1.77	0.86	-14.98	0.22	-12.21	3.46
1-person households (n = 165)	-12.43	1.63	0.96	-13.84	0.23	-11.02	3.94
2-person households (n = 333)	-13.99	1.84	0.99	-15.29	0.27	-12.69	3.64
3+ person households (n = 188)	-12.79	1.67	0.91	-14.24	0.21	-11.34	3.86
(by % shopping with retailer)							
Low-medium loyalty (0-60%) (n = 244)	-13.27	1.72	0.82	-14.65	0.21	-11.89	3.25
High loyalty (61-80%) (n = 174)	-13.11	1.72	0.96	-14.46	0.25	-11.77	3.70
Very high loyalty (81-100%) (n = 268)	-12.49	1.66	1.13	-13.83	0.30	-11.16	4.30

A = average of purchased energy and individual energy intake. For the purposes of comparison, all values are presented for A = 2000 kcal.

5.3.4 Agreement for relative estimates

Energy-adjusted estimates give an impression of the relative composition of the diet, regardless of volumes purchased or consumed. For relative dietary composition, the two measures (purchase and intake) are in much closer agreement than was observed for absolute values (Table 5.7), as evidenced by a lesser gradient of regression lines, and closer limits of agreement (Figure 5.4). To aid comparison between nutrients, all results presented in Table 5.7 are stated for the value of A (average of measures) at which the difference is zero (ratio of difference = 1) For example, where sugar makes up 26.8% of total energy on average across purchases and intake, the mean of the difference is zero.

While difference and directionality remain related to magnitude, this is to a lesser degree. The closest agreements are observed for sugar (where the ratio of difference = 1, LoA 0.59 – 1.67) and saturated fat (LoA 0.62 – 1.60). The greatest difference in agreement is observed for sodium, for which where the ratio of difference = 1, purchases are likely to under-estimate sodium/kcal intake by up to a half, or over-estimate it by up to two times.

Figure 5.4a) shows that purchases estimates of the proportion of total energy to which sugar contributes are typically lower than intake estimates below a mean value of around 25%. Purchases typically estimate a lower proportion from protein, total fat and saturated fat up to a mean value of around 20%, 36% and 12.5% respectively. Below a mean of around 1.4mg/kcal, estimates of sodium (mg) per kcal purchased are typically lower than estimated sodium (mg) per kcal consumed (Figure 5.4e).

Table 5.7 Regression coefficients for difference and limits of agreement for energy-adjusted purchase and intake for macronutrients and sodium (whole sample, n=686)

	Mean difference (purchase / intake)			**Lower limit of agreement**		**Upper limit of agreement**	
	Intercept (b_0)	**Slope (*b1*)**	**A (for ratio of difference = 1)**	**Intercept (b_0)**	**Ratio of difference**	**Intercept (b_0)**	**Ratio of difference**
Sugar (% energy)	-1.25	0.38	26.8	-1.77	0.59	-0.74	1.67
Protein (% energy)	-2.47	0.84	18.9	-3.05	0.56	-1.90	1.77
Total fat (% energy)	-3.04	0.84	37.3	-3.51	0.62	-2.57	1.60
Saturated fat (% energy)	-1.93	0.73	14.1	-2.58	0.52	-1.28	1.92
Sodium (mg/kcal)	-0.46	1.35	1.4	-1.16	0.50	0.23	1.97

A = average of purchased energy and individual energy intake. For the purposes of comparison, all values are presented for the value of A at which the ratio of the difference = 1 (no difference between measures).

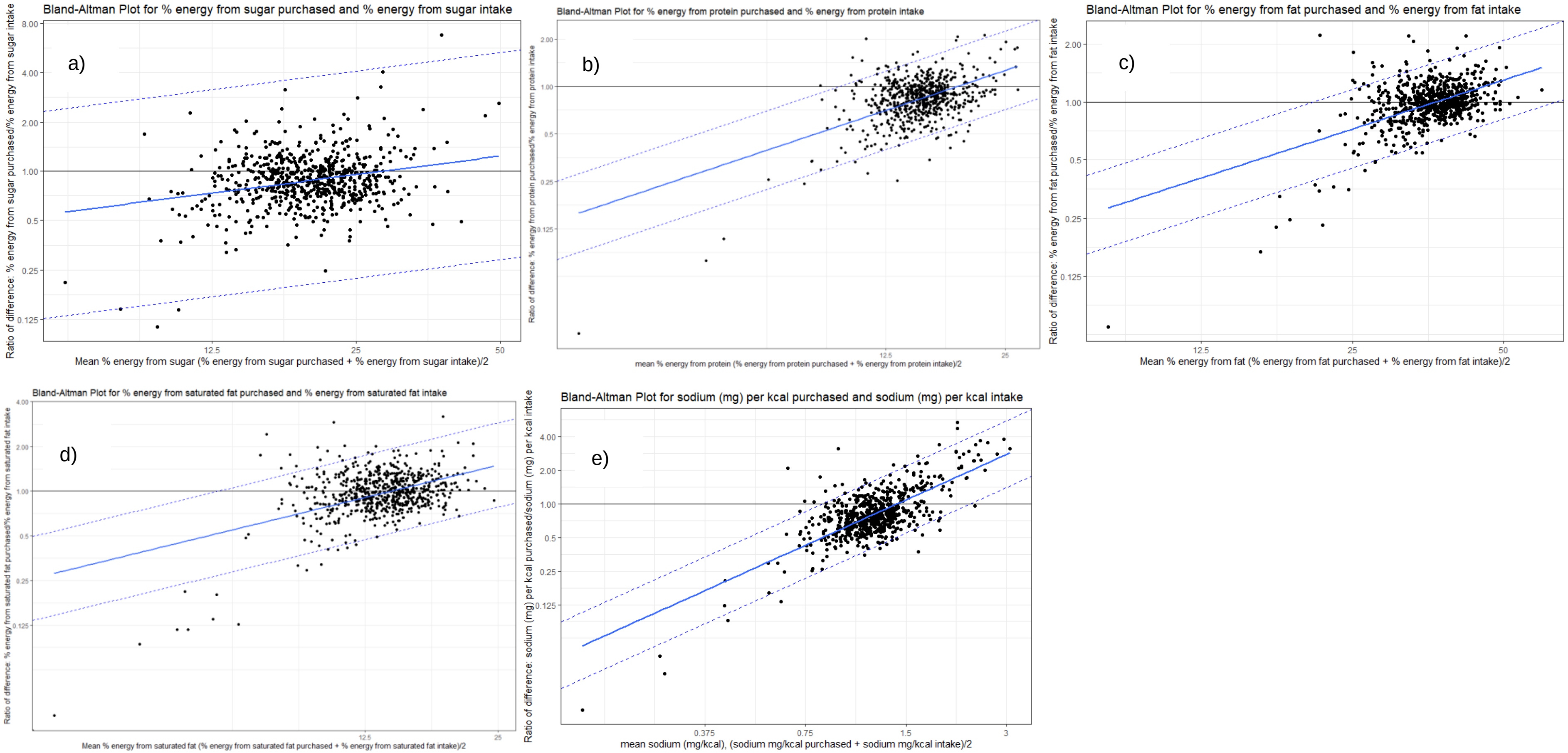


Figure 5.4 Bland-Altman plots for ratio of energy-adjusted nutrient purchased (individual-level)/energy-adjusted nutrient intake, plotted against their average, by nutrient.

a) Sugar, b) Protein, c) Total fat, d) Saturated fat, e) Sodium

5.4 Discussion

This paper assesses the agreement between daily intake estimates and daily loyalty card purchase estimates, for energy and five key nutrients (sugar, protein, total fat, saturated fat, and sodium). Using a unique study dataset, the STRIDE study found agreement to be strongest for smaller households and among the most loyal customers. Absolute purchase values (be they at the household or individual-level) were found to be a poor proxy for individual intake, yet purchases represented dietary composition relatively well making them a good marker of dietary intake pattern. By nutrient, the strongest agreements were found for sugar and saturated fat, and for energy-adjusted values in particular. The STRIDE study contributes to evidence for the validity of purchase records as a proxy for dietary intake. To our knowledge, this is the first study to quantify the statistical agreement and limits to agreement between actual and energy-adjusted nutrient estimates from automated electronic purchase data and self-reported intake.

Electronically captured purchase records have appeal for their use in population dietary assessment due to their scalability and automated nature. While an obvious limitation is that we do not know exactly what proportion of each customer’s food purchases were carried out with the retailer, we have accounted for some of this variability by asking participants to self-report the retailer’s contribution to their shopping. In addition, it is possible that not all purchases at the retailer may be captured by the data, if they forget to scan their loyalty card for example. That said, automated collection reduces participant and researcher burden and limits the chance of purchases being consciously or sub-consciously affected by participation in the study.

Previous comparison studies have described purchase data as a moderately good indicator of intake (Vepsalainen et al., 2021; Eyles et al., 2010) according to correlation of nutrient amounts (Eyles et al., 2010) and association by food category volume and frequency (Vepsalainen et al., 2021). Despite this conclusion, the comparison methods applied were unable to estimate the magnitude of the agreement. As a result, adjustment factors were unavailable to allow for conversion between methods, until now. Using the Bland-Altman method for quantifying statistical agreement, which is considered the gold-standard comparator for validation of health research methods (Bland and Altman, 1986; Bland and Altman, 1999), this study provides a starting point towards developing such adjustment factors.

This study found overall, household purchase estimates to be a poor proxy for intake estimates, across all nutrients. Limits of agreement were wide and agreement was also found to be related to magnitude of the mean of estimates. At greater magnitudes, purchase estimates were several times higher than reported intake, even after extrapolation of purchases to the individual-level, while at lower magnitudes purchase estimates represent just a small fraction of total intake

It is likely that over-estimation is due to a combination of; large household sizes and inaccuracies in our individual purchase proxy, food waste which may be particularly high among some customers, purchasing for other households (13% of respondents to the STRIDE baseline questionnaire reported purchasing for others outside of the household as a change in their shopping habits since the COVID-19 pandemic began), and a systematic under-reporting of intake by some participants. While average food waste is estimated to be around 10% (Wrap, 2018), self-reported intake is thought to underestimate true energy consumption by a similar degree (Ravelli and Schoeller, 2020), thus it is possible that these errors cancel out. Where purchases under-estimate intake this is most likely due to foods purchased elsewhere and thus not captured by supermarket loyalty card transactions for a single retailer.

Our study is unique in that we attempted to account for household composition to calculate an individual-level purchase estimate for participants. After extrapolation to the individual level, purchase data became more likely to under-estimate intake. Subgroup analysis showed that agreement with individual-level purchase estimates was poorer for larger households. We expect this is due to error built in by the method for allocating nutrients to household members, which increases as the number of people in the household increases. By allocating all nutrients in accordance with age-specific energy intake recommendations, regardless of their food source (for example, energy derived from alcohol is allocated to children as well as adults), we anticipate greater error for households containing children. The contribution of school meals to children’s diets, differing ratios for dietary requirements by nutrient (e.g. low salt diets recommended for infants), and unknown genders of other household members, could constitute further sources of error. Despite poor agreement between purchases and intake for absolute nutrient values, our findings mirrored those from Vepsalainen et al. (2021), in that we found closer agreement for smaller households and more loyal customers (according to self-reported proportion of shopping with the study retailer).

Differences in agreement by nutrient were observed, in line with previous findings which reported strongest relationships for total fat and saturated fat (Eyles et al., 2010; Ransley et al., 2001) and variation in concurrence by food group (Vepsalainen et al. 2021). Agreement for saturated fat was slightly lower, which may be due to a higher tendency to purchase high-saturated fat treat items elsewhere. For example, crisps and sweet treats are often consumed on the go, purchased at cafes, petrol stations and from vending machines (Einav, 2008). The lowest agreement was observed for sodium, which may reflect that other food sources (e.g. out of home) contribute a relatively higher proportion of salt to the diet (restaurant and takeaway meals have been found to contain higher levels of salt than home-cooked or ready meal equivalents (PHE, 2018)). Or, purchase data may poorly account for table salt added to food at home, which tends to be purchased in large quantities but relatively infrequently. It is also likely that the time period covered by purchase data influences the degree of agreement with consumption of salt and other store-cupboard items. This theory is supported by findings by Vepsalainen et al. (2021), who reported weak associations for vegetable oil, as well as a general trend for stronger associations when comparing intake with 12-months purchase data compared with just one month. Exploration of the timescale required of purchase records to capture habitual diet is therefore warranted.

Adjusting nutrients for total energy allows for comparison of relative dietary composition, rather than absolute nutrient quantities. As expected, energy-adjusted nutrient purchases showed a higher agreement with intake, particularly for total fat and sugar, similar to the observations made by Eyles et al. (2010) who also reported a high correlation for total fat. Furthermore, the relationship with customer loyalty across energy-adjusted measures was less apparent than for absolute measures. This indicates that while purchases from a single retailer tend to under-estimate nutrient intake in absolute terms, they are relatively reflective of overall dietary choices. Proportion of purchased energy from macronutrients could therefore provide a useful surrogate marker for dietary quality (Appelhans et al., 2017). This supports the validity of transaction data in dietary patterns research (Clark et al., 2021) and for ecological research applications, such as evaluating policy impacts (forthcoming (Jenneson et al., 2022)) and identifying population-level trends such as the increasing popularity of plant-based protein sources (Piernas et al., 2021).

A limitation of this study is that, due to its prospective nature, it was not possible to sample customers based on their loyalty to the retailer during the study period. While we made an attempt to account for customer loyalty by selecting customers who purchased regularly with the supermarket during the year prior to recruitment, it was apparent that previous loyalty did not reflect customer purchasing behaviours during the study period. This observation may have been unique due to the circumstances of the COVID-19 pandemic, which saw many customers switching to different retailers due to the proximity of stores or availability of online delivery slots. Studies using retrospective sampling approaches should therefore account for customer loyalty when selecting study participants to improve representation of intake. If customer cohorts are to be followed over time, characterisation of customer loyalty and re-sampling are likely to be beneficial to ensure the sample remains representative of loyal customers.

A strength of this study is the use of a bespoke product nutrient composition database which combines product-level composition data from the back of pack nutrition label with generic food composition data. By using actual product composition information where possible, the accuracy of nutrient estimates from purchase data in maximised. Indeed, it may be true that for some foods (particularly for composite dishes, for which there may be just one option available in generic food tables compared with many different products on the retailer's shelves), nutrient estimates at the product level are likely to more accurate than for intake estimates as they enable accounting for brand-level differences. Furthermore, the database used gave extremely good coverage of product nutrient data across all food categories, rather than being restricted to just the most commonly purchased foods as in the study by (Eyles et al., 2010).

5.4.1 Future research avenues

The STRIDE study provides a rich dataset which will enable further investigation of differences in agreement according to customer demographic characteristics (such as age, gender, and BMI), by season, and by geography (according to geodemographic classification and are-level deprivation indices).

Future work should also explore what volume of transaction records are most suitable for assessing habitual diets, taking into account their ability to capture less frequently purchased bulk or store-cupboard items, and seasonal dietary patterns. Additionally, methods are required to estimate household size and composition, in the absence of survey data.

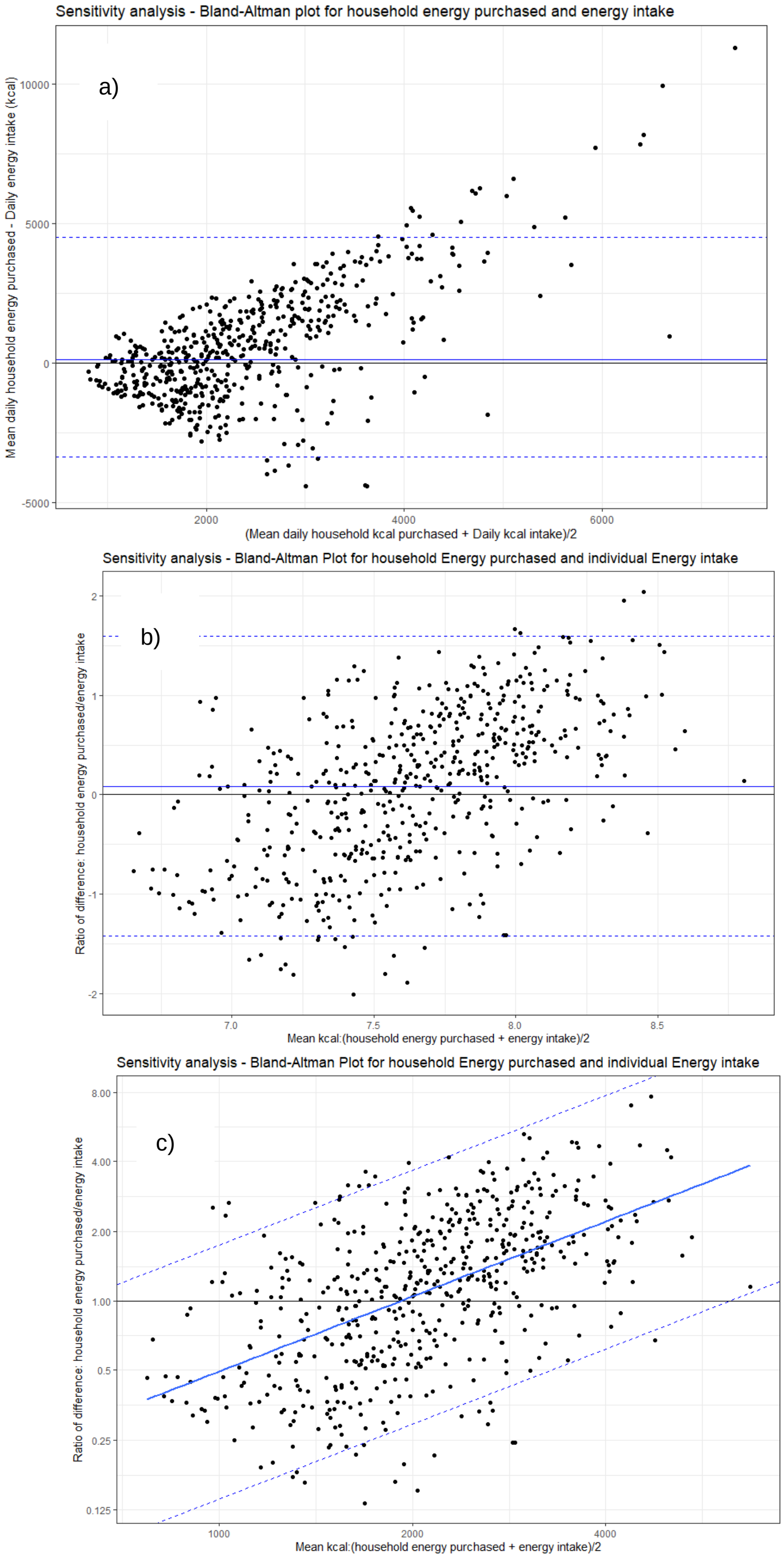
5.5 Conclusion

This study demonstrates progress towards the generation of adjustment factors for extrapolation of household purchase estimates to individual intake estimates. In this setting, where it was not possible to restrict the customer sample to only shoppers who purchase most of their food from the study supermarket, we found poor agreement between absolute nutrient measures from purchase data and self-reported intake. Agreement was strongest for single-person households, loyal customers, energy, total fat, and sugar, providing evidence that customer sampling is an important consideration for studies using supermarket transaction data. Energy-adjusted nutrient estimates provide a good indicator of dietary composition (which appears to be unrelated to customer loyalty), which may be beneficial for ecological studies, identification of intervention target groups, and monitoring of dietary patterns and quality with applicability for policy evaluation.

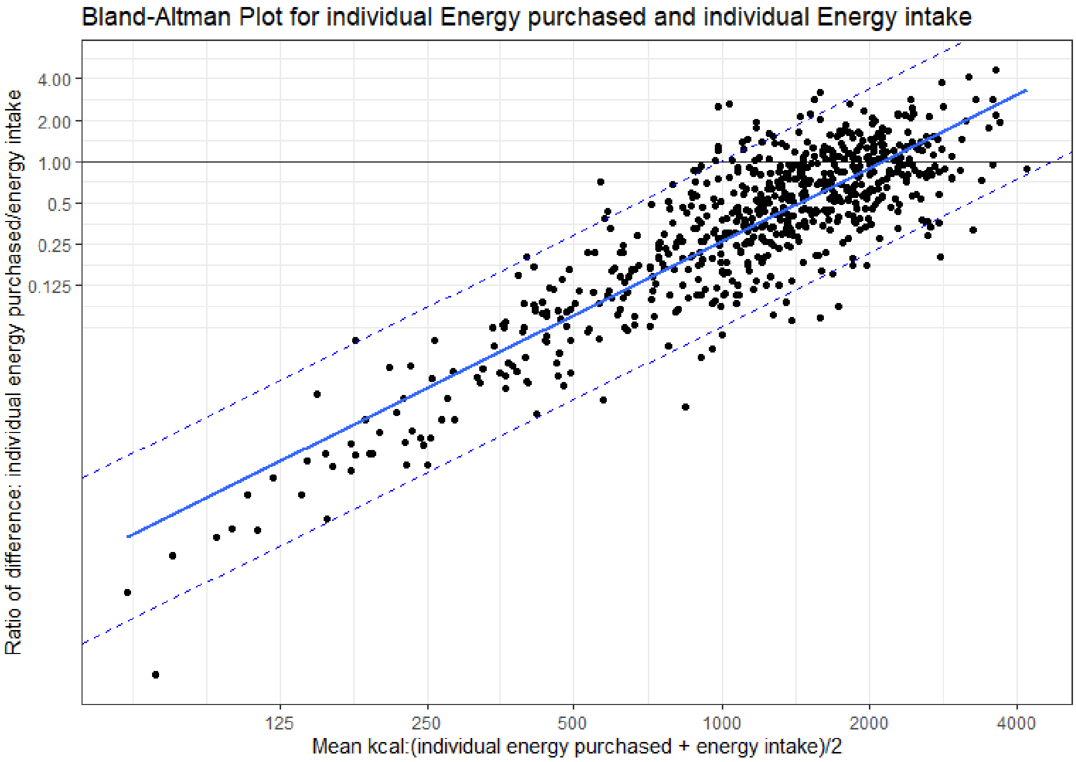
5.6 Supplementary materials

Supplementary Table 1. Mean difference and limits of agreement between purchase and intake for energy (kcal)

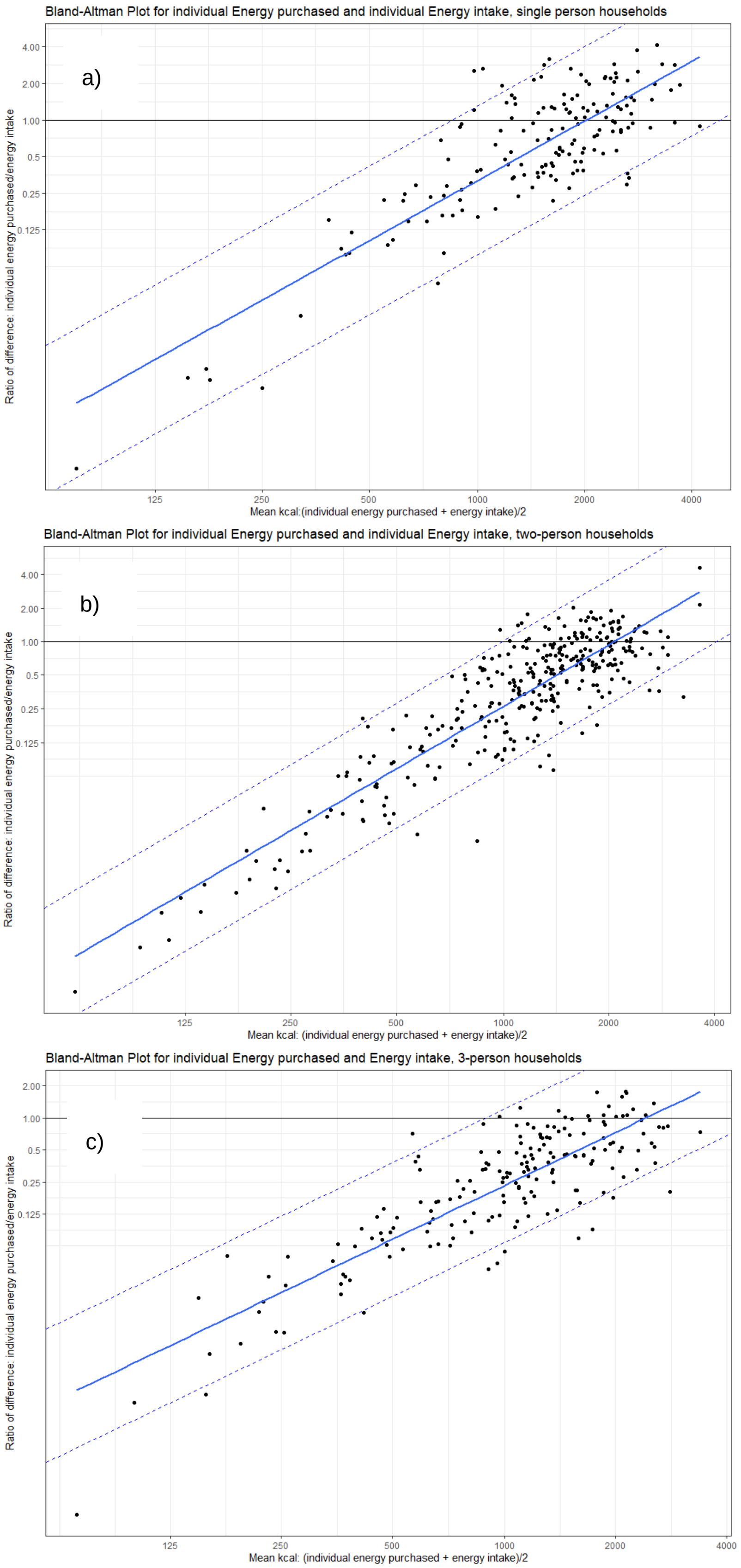
	Mean difference, (purchase – intake) kcal (SD)	Difference (purchase/intake) expressed as a percentage (%)	Lower limit of agreement (%)	Upper limit of agreement (%)
Household purchase – Intake				
All households	129 (2067)	66	5	987
Individual purchase – Intake (by household size)				
All households	-971 (1328)	33	2	531
1-person households (n = 165)	-329 (1495)	58	5	701
2-person households (n = 333)	-1027 (1207)	31	2	505
3+ person households (n = 188)	-1436 (1153)	22	1	329
Individual purchase – Intake (by % shopping with retailer)				
Low-medium loyalty (0-60%) (n = 244)	-1533 (1031)	19	1	272
High loyalty (61-80%) (n = 174)	-916 (1189)	25	2	587
Very high loyalty (81-100%) (n = 268)	-495 (1458)	51	4	675



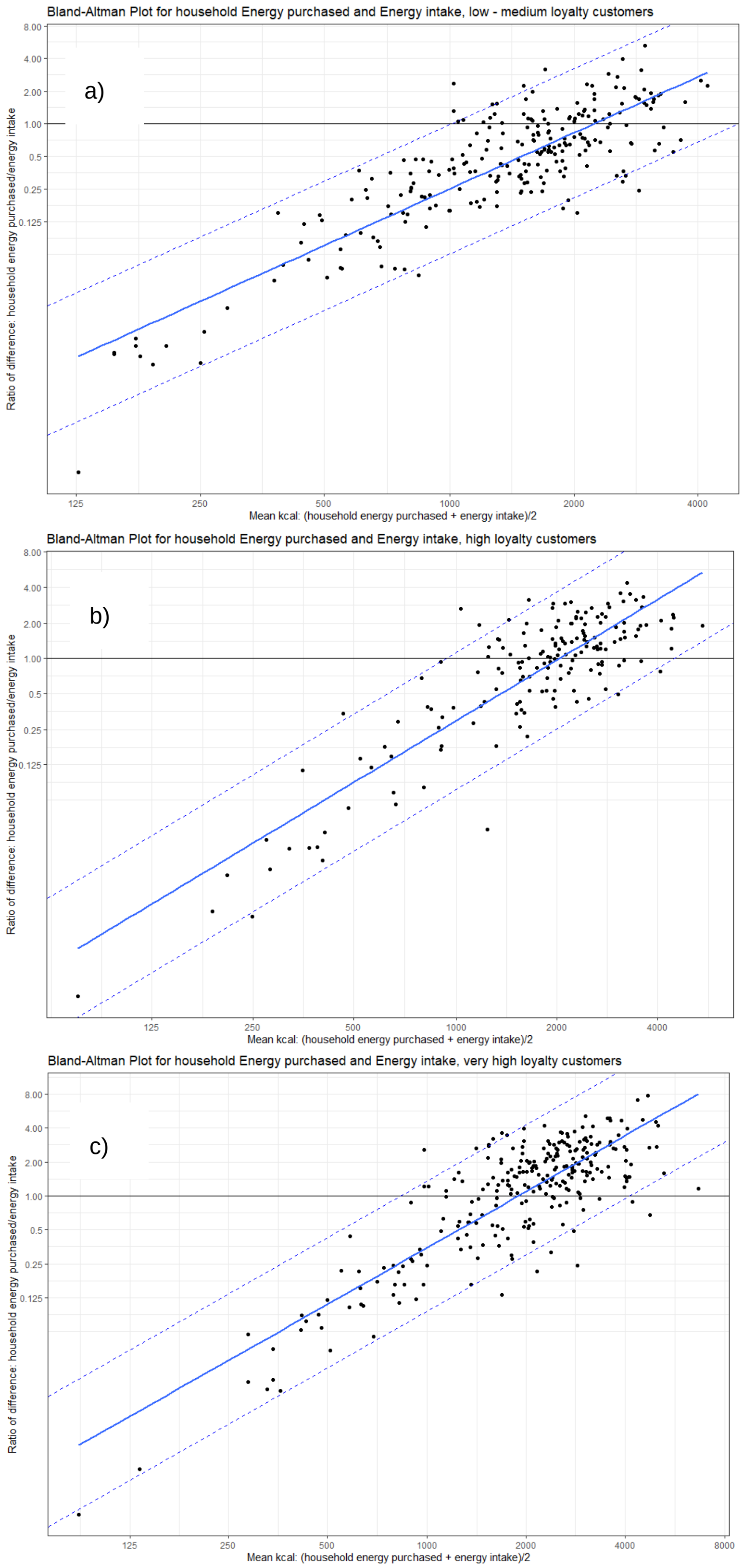
Supplementary Figure 1. Bland-Altman plots showing the agreement and limits of agreement between daily household energy purchased and daily energy intake (kcal), sensitivity analysis excluding customers purchasing <500kcal/day on average (n=562)



Supplementary Figure 2. Bland-Altman plots showing the agreement and limits of agreement between daily individual energy purchased and daily energy intake (kcal)



Supplementary Figure 3. Bland-Altman plots showing the agreement and limits of agreement between daily individual energy purchased and daily energy intake (kcal) by household size; a) single-person households, b) 2-person households, c) 3+ person households



Supplementary Figure 4. Bland-Altman plots showing the agreement and limits of agreement between daily individual energy purchased and daily energy intake (kcal) by customer loyalty; a) low-medium loyalty, b) high loyalty, c) very high loyalty

References

Appelhans, B.M., French, S.A., Tangney, C.C., Powell, L.M. and Wang, Y. 2017. To what extent do food purchases reflect shoppers’ diet quality and nutrient intake? *International Journal of Behavioral Nutrition and Physical Activity.* **14**(1), p46.

Bland, M.J. 2005. *The Half-Normal distribution method for measurement error: two case studies*. York, UK: University of York. [Talk first presented in the Statistics and Econometrics seminar series, University of York, February 2005]. Available from: https://www-users.york.ac.uk/~mb55/talks/halfnor.pdf

Bland, M.J., Altman, D.G. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* **i**, pp.307-310.

Bland, M.J., Altman, D.G. 1999. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research.* **8**, pp.135-160.

Burley, V., Timmins, K., Cade, J., Greenwood, D., Husain, F., Gill, V., Vowden, K., Page, P., Lennox, A., Erens, B., Steer, T. and Hulme, C. 2015. *Making the best use of new technologies in the National Diet and Nutrition Survey: a review.*

Buttriss, J.L., Welch, A.A., Kearney, J.M. and Lanham-New, S.A. 2017. *Public Health Nutrition.* Wiley.

Carter, M.C., Hancock, Neil., Albar, Salwa A., Brown, Helen., Greenwood, Darren C., Hardie, Laura J., Frost, Gary S., Wark, Petra, A., and Cade, Janet E. 2016. Development of a New Branded UK Food Composition Database for an Online Dietary Assessment Tool. *Nutrients.* **8**(480).

Clark, S.D., Shute, B., Jenneson, V., Rains, T., Birkin, M. and Morris, M.A. 2021. Dietary Patterns Derived from UK Supermarket Transaction Data with Nutrient and Socioeconomic Profiles. *Nutrients.* **13**(5), p1481.

de la Hunty, A., Buttriss, J., Draper, J., Roche, H., Levey, G., Florescu, A., Penfold, N. and Frost, G. 2021. UK Nutrition Research Partnership (NRP) workshop: Forum on advancing dietary intake assessment. *Nutrition Bulletin.* **46**(2), pp.228-237.

Department of Health. 2017. *Technical guidance on nutrition labelling.* London, UK: UK Government.

Einav, L., Leibtag, Ephraim., Nevo, Aviv. 2008. *On the Accuracy of Nielsen Homescan Data.* U.S. Department of Agriculture.

Eyles, H., Jiang, Y. and Ni Mhurchu, C. 2010. Use of Household Supermarket Sales Data to Estimate Nutrient Intakes: A Comparison with Repeat 24-Hour Dietary Recalls. *Journal of the American Dietetic Association.* **110**(1), pp.106-110.

Gale, C., Singleton, A., Bates, A. and Longley, P. 2016. Creating the 2011 area classification for output areas (2011 OAC). *Journal of Spatial Information Science.* **12**.

GOV.UK. 2015. *English indices of deprivation 2015.* [Online]. [Accessed 12.01.2022]. Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015

GOV.UK. 2020. *Collection: Family food statistics. Annual statistics about food and drink purchases in the UK.* [Online]. [Accessed 12.01.2022]. Available from: https://www.gov.uk/government/collections/family-food-statistics

Green, M.A., Watson, A.W., Brunstrom, J.M., Corfe, B.M., Johnstone, A.M., Williams, E.A. and Stevenson, E. 2020. Comparing supermarket loyalty card data with traditional diet survey data for understanding how protein is purchased and consumed in older adults for the UK, 2014–16. *Nutrition Journal.* **19**(1), p83.

Jenneson, V., Dyer, J., Matousek, A., Francois, I., Tumelty, J., Omieljaniuk, M., Stephens, M., Martin, R., Wu, S., Yung Low, S., Yip, W. and Morris, M.A. 2022. *Investigating the impact of the UK's Soft Drinks Industry Levy on consumers' purchases of soft drinks.* London, UK: The Alan Turing Institute Data Study Group.

Jenneson, V., Morris, M., Greenwood, D. and Clarke, G. 2020. *STRIDE Study (Supermarket Transaction Records In Dietary Evaluation) protocol*. Open Science Framework. Available from: https://doi.org/10.17605/OSF.IO/VUKTQ

Jenneson, V., Pontin, F., Greenwood, D.C., Clarke, G.P. and Morris, M.A. 2021. A systematic review of supermarket automated electronic sales data for population dietary surveillance. *Nutrition Reviews.*

Masson, L.F., McNeill, G., Tomany, J.O., Simpson, J.A., Peace, H.S., Wei, L., Grubb, D.A. and Bolton-Smith, C. 2003. Statistical approaches for assessing the relative validity of a food-frequency questionnaire: use of correlation coefficients and the kappa statistic. *Public Health Nutrition.* **6**(3), pp.313-321.

Nevalainen, J., Erkkola, M., Saarijärvi, H., Näppilä, T. and Fogelholm, M. 2018. Large-scale loyalty card data in health research. *Digital health.* **4**, pp.2055207618816898-2055207618816898.

PHE. 2016. *Government Dietary Recommendations. Government recommendations for energy and nutrients for males and females aged 1 - 18 years and 19+ years.* London, UK: UK Government.

PHE. 2018. *Salt targets 2017: Progress report. A report on the food industry's progress towards meeting the 2017 salt targets.* London, UK.

PHE. 2020. McCance and Widdowson's composition of foods integrated dataset. London, UK: *gov.uk*. [Online]. [Accessed 21.01.2022]. Available from: https://www.gov.uk/government/publications/composition-of-foods-integrated-dataset-cofid

Piernas, C., Cook, B., Stevens, R., Stewart, C., Hollowell, J., Scarborough, P. and Jebb, S.A. 2021. Estimating the effect of moving meat-free products to the meat aisle on sales of meat and meat-free products: A non-randomised controlled intervention study in a large UK supermarket chain. *PLOS Medicine.* **18**(7), pe1003715.

Public Health England. 2019. *NDNS: time trend and income analyses for Years 1 to 9.* Public Health England., and the Food Standards Agency.,.

Ransley, J.K., Donnelly, J.K., Khara, T.N., Botham, H., Arnot, H., Greenwood, D.C. and Cade, J.E. 2001. The use of supermarket till receipts to determine the fat and energy intake in a UK population. *Public Health Nutrition.* **4**(6), pp.1279-1286.

Ravelli, M.N. and Schoeller, D.A. 2020. Traditional Self-Reported Dietary Instruments Are Prone to Inaccuracies and New Approaches Are Needed. *Frontiers in Nutrition.* **7**.

Simpson, E., Bradley, J., Poliakov, I., Jackson, D., Olivier, P., Adamson, A.J. and Foster, E. 2017. Iterative Development of an Online Dietary Recall Tool: INTAKE24. *Nutrients.* **9**(2), p118.

Vepsalainen, H., Nevalainen, J., Kinnunen, S., Itkonen, S.T., Meinila, J., Mannisto, S., Uusitalo, L., Fogelholm, M. and Erkkola, M. 2021. Do we eat what we buy? Relative validity of grocery purchase data as an indicator of food consumption in the LoCard study. *Br J Nutr.* pp.1-24.

Wrap. 2018. *Household food waste: restated data for 2007-2015.*

Part 3

Discussion and Conclusions

Chapter 6
Discussion

6.1 Overview

The field of dietary assessment has long sought inexpensive, scalable, and objective population dietary surveillance methods. Recently, attention has turned to technology to supplement traditional paper-based self-report methods. Interest in the use of supermarket transaction records as a dietary indicator is growing, yet few studies to date have attempted to understand just how well they represent dietary intake. Linking big data analytics and spatial exploratory techniques, together with dietary assessment, the work in this thesis demonstrates how an interdisciplinary approach adds value to the application of supermarket transaction records for population dietary monitoring. Furthermore, as the first study to quantify agreement between household transactions and individual nutrient intake, the STRIDE study for the first time enables conversion between measures.

This thesis set out with the aim:

To understand and evaluate the contribution that supermarket loyalty card transaction records can make to population dietary monitoring, both on a national and a small-area scale.

The preceding chapters describe the exploratory approach undertaken to achieve this aim, and present the findings in the form of three papers:

Paper 1 (Chapter 3) – *A systematic review of automated electronic supermarket sales data for population dietary surveillance.*
Paper 2 (Chapter 4) - *Exploring the geographic variation in fruit and vegetable purchasing behaviour using supermarket transaction data.*
Paper 3 (Chapter 5) - *A validation study: Supermarket Transaction Records In Dietary Evaluation (STRIDE).*

In summary, my literature synthesis in Chapter 3 (Jenneson et al., 2021) found supermarket transactions a useful population-level metric for: dietary surveillance; policy evaluation; and assessing the success of interventions, particularly those which alter the food environment. In Chapter 4 (Jenneson, et al., 2022), my unique application of loyalty card transactions to geographic dietary exploration revealed spatial clustering in dietary behaviours at the neighbourhood level. Through linkage with other data sources, transaction records have the capacity to contribute a better understanding of the relationship between our food environment and dietary behaviours, and the role of deprivation and cultural moderators. This has applicability for ecological research, hypothesis generation and the development and monitoring of place-based interventions and policy.

The STRIDE study (presented in Chapter 5) is the first of its kind to directly quantify the agreement between purchase and intake for absolute and energy-adjusted nutrient metrics, at both the household and individual-levels. My findings contribute novel understandings of the validity of transaction records as a dietary assessment tool. In the context of the STRIDE study, where it was not possible to sample customers based on their loyalty to the supermarket, absolute amounts of purchased nutrients were found to be a poor proxy for intake. However, comparatively good agreement was observed for energy-adjusted nutrient metrics, suggesting supermarket transactions provide a reasonable indication of dietary composition. I found agreement to vary by magnitude, household size and retailer loyalty, indicating that transaction records are most indicative of intake among smaller households containing loyal customers. By presenting the average agreement and limits of agreement as regression equations, the reader can estimate the expected agreement at a given magnitude. This study therefore makes an important contribution to the development of adjustment factors for conversion between household purchases and individual intake.

The key findings from this thesis are listed below:

6.1.1 Key Findings

1. Transactions have applicability for population dietary surveillance, intervention assessment and policy evaluation
2. Loyalty cards enable longitudinal follow-up of household dietary trends without burden to the data subject
3. Large sample sizes are possible, with good coverage even in hard-to-reach low-income groups
4. Their geocoded nature makes applications to local interventions and policy plausible
5. Dietary purchase behaviours cluster spatially and reveal small-area patterns in purchasing
6. Agreement with individual intake varies according to household size and customer loyalty
7. Household purchases offer a good indicator of dietary composition, useful for assessing dietary quality and dietary patterns
8. Loyal customer samples with ‘complete’ purchases represented by the retailer are required to gain reliable dietary estimates from single retailer data.

In this final chapter, I draw together the evidence from each of the three papers, critically evaluating the findings in the wider research context and in line with the theoretical Framework for Evaluating Diet (FED) introduced in Chapter 1. I demonstrate how each of the nine objectives introduced in Chapter 1 were met, in order to achieve the thesis' aim.

6.2 Nutrition data metrics

There exist a number of metrics through which diet may be measured: at the nutrient-level; the food group-level; or a combination of both. As all metrics have their strengths and limitations, the choice of dietary metric is dependent upon the use case or research question to be answered (WHO, 2021). For example, nutrient adequacy metrics are useful for assessing risk of over-or under-nutrition; food-group metrics are important for assessing alignment with food-based dietary guidelines (such as comparison against the Eatwell guide (NHS, 2019), or measuring fruit and vegetable consumption against the UK's 5-a-day recommendation (NHS, 2018)); and diet-quality metrics such as the Healthy Eating Index (USDA, 2020) and Diet Quality Score (Toft et al., 2007) support the identification of groups at risk of poor diet-related health outcomes. Metrics reported in nutrition research lack consistency across time and place, making between-study and global dietary comparisons challenging (WHO, 2021). There is thus a need for robust metrics which are simple, feasible, scalable, and allow for comparability over time and space (WHO, 2021).

As demonstrated by their position in the FED diagram (Chapter 1, Figure 1.1), purchase data (Stage 2) capture an important interaction between people and their food environment (Stage 1). Food purchasing represents the first opportunity for consumers to exercise dietary choice, albeit within the constraints of food availability, financial circumstance, and their own knowledge, values and beliefs. By capturing these purchase moments, transactions offer insight into dietary trends over time; for example, how the sales of certain products fluctuate in response to interventions, changes in economic situation, or evolving cultural narratives (Jenneson et al., 2021). In addition, transactions offer a useful and convenient tool for providing immediate dietary feedback at the customer-level. While their application to population dietary monitoring is the focus of this thesis, it is of policy relevance to recognise their potential for application in interventions to encourage behaviour change, when presented in the context of dietary quality of individual food choices (e.g. to encourage healthy swaps) or at the basket level (An et al., 2013; An and Sturm, 2017; Piernas, et al., 2019; Piernas, et al., 2020). In my systematic review, I found the popularity of supermarket transaction records an inexpensive, large-scale objective dietary research tool to have grown over recent years (Jenneson et al., 2021). Their automated nature, continuous longitudinal collection and national coverage offer a potential solution for population-level dietary monitoring. In this section, I consider how well-suited purchase records are for reporting across a range of dietary metrics and discuss their applicability for use at a national and global level.

6.2.1 Product coverage

Supermarket transaction records permit nutrient-level dietary monitoring through linkage to product nutrient composition data via a unique product identification number, such as the SKU (Stock-Keeping Unit), EAN (European Article Number) or GTIN (Global Trade Item Number). I demonstrate how this may be achieved in the STRIDE study presented in Chapter 5 of this thesis. Yet, in my synthesis of the literature (Chapter 3) I found relatively few studies using transaction records to report dietary outputs at the nutrient-level (Jenneson et al., 2021). This is due in large part to the lack of readily available product-level nutrient information for research, to which private companies charge for access.

Even with the benefit of data access, the instability of product identifiers is problematic to the theoretical ease of data linkage. The high-frequency of product turnover on the market means that static product composition data cuts (such as the one used in the STRIDE study) do not offer a full coverage for transactions gathered longitudinally over several months. Un-matched products tended to be seasonal items such as Easter eggs and Christmas treat foods, which vary year on year in response to market trends. It is possible that this may translate to seasonal variation in agreement with self-report intake measures, but this is yet to be explored. Other items which are likely to be missed are new entrants onto the market and reformulated products where the change in composition is substantial enough to warrant a new product listing. It is therefore important that obtained product data be up to date and compatible with the transaction period.

Another limitation of linkage with product composition data is that it only covers pre-packaged goods for which nutrition labelling is required. UK labelling regulations do not require product composition data to be available for unpackaged items (e.g. fresh fruits and vegetables, and deli-counter products such as cheese, cooked meats and antipasti) and alcohol, which were all missing. While previous studies have excluded alcohol, I chose to include it in recognition of its significant contribution to energy intake (Fong et al., 2021). In the UK, there has been an increase in consumption of alcohol at home (Foster and Ferguson, 2012), possibly accelerated by the introduction of strict drink-driving laws and the closure of pubs. As a result, supermarkets have become a major source of alcohol (IAS, 2018). This trend is further compounded during the study time-period by the COVID-19 pandemic and lockdown restrictions which saw the closure of pubs, bars and restaurants (Institute for Government., 2021).

To fill in the gaps due to uncaptured products, supplementation of product data with generic values in national dietary composition tables is necessary. This approach was taken in the validation study by Eyles et al., (2010), however, the combined nutrient dataset was limited to the 3,000 top-selling products (by volume) (Hamilton et al., 2007), representing less than 20% of the retailer product list. While pragmatic, this may introduce bias in the agreement across customers, particularly for minority ethnic and low-income groups who are more likely to purchase specialist ethnic and retail own-brand foods which have a lower product turnover. To avoid potential bias due to product selection, I mapped each un-matched food and beverage product purchased by STRIDE customers to the closest matching item in UK CoFID tables (PHE, 2020b), achieving an excellent (100%) match-rate. This was a manual process due to the absence of common food codes, and relied upon keyword searches and my expertise as a nutritionist. While product coverage was a strength for the STRIDE study, the resource requirements for mapping are potentially prohibitive to repeatability to other timepoints and settings, suggesting the need for alternative solutions.

My aim in developing the bespoke food composition database for STRIDE was to represent the nutrition composition of purchased foods as accurately as possible. Product-level accuracy is a strength of purchase data which can reveal brand-level differences in response to policy, and customer-level differences in brand loyalty which may translate to differences in dietary composition by sub-group and over time (forthcoming, (Jenneson, et al., 2022)). Yet accuracy varies by food group due to the reliance on national dietary tables, for produce, in-store bakery items, and alcohol, for example. Furthermore, the different sources of nutrient composition data used for intake and purchase metrics may contribute to some of the observed difference in agreement, though the extent of this is unknown. Discrepancies are likely to be greater for ethnic dishes and new product innovations, such as plant-based dairy alternatives and meat substitutes (e.g. Jackfruit) which are not well covered in national composition databases. Comparison with intake reported through the myfood24 tool (Carter et al., 2015), which maps consumed foods to a product-level nutrient database containing around 45,000 branded and retailer own-brand products, may eliminate some of this difference.

To fulfil the potential of purchases as a stable global dietary metric at the nutrient-level, practical issues of data acquisition and linkage must first be overcome. Web-scraping methods are a useful method for gathering up-to-date product nutritional information (Chidambaram et al., 2013; Harrington et al., 2019), where it cannot be obtained from the retailer. Yet, product-level data will probably always require supplementation with national food tables for uncaptured products. Linkage to national food composition tables alone may be achievable in the absence of product-level data, but accuracy is limited to just 2-3000 products typically, limiting the utility for monitoring reformulation effects and the impact of brand choice on dietary quality. Furthermore, linkage would require automated mapping approaches such as natural language processing, in the absence of common product identification keys (Tran et al., 2017; Chidambaram et al., 2013; Carter et al., 2016).

6.2.2 Nutrient coverage and accuracy

The product-level nutrition data reported reflects local packaging requirements. While this is advantageous for standardisation of metrics within a region, the coverage of nutrients is limited and varies globally. In the UK, mandatory back of pack nutrient reporting covers only energy, carbohydrates, sugars, total fat, saturated fat, protein, and sodium (Department of Health., 2017). It is not mandatary to report the fibre content of all foods (Department of Health., 2017), making coverage of fibre data across the market incomplete. This is fairly typical of product labelling requirements globally, limiting the capacity for transaction data as a metric for micronutrient quantity.

Some variation also exists in how on-pack nutritional information may be expressed, which has implications for accuracy of nutrient estimates from purchase data. UK guidance states that nutrition information should be stated per 100g of product as sold, unless preparation guidelines are provided (Department of Health., 2017). For example, in the case of instant noodles or fruit cordials where the amount of water to be added is stated on the pack, the nutritional value may be expressed for the product as consumed. However, in the absence of full on-pack instructions within nutrition data cuts, it is not possible to identify whether values are expressed as sold or as consumed. A further potential source of inaccuracy is the provision of nutrition data for some products by weight (per 100g of product) and for others by volume (per 100ml of product), without a specific gravity value to enable conversion to a common metric. Furthermore, specific gravity data in CoFID is sparsely reported (less than 2% of products) (PHE, 2020b). Due to a lack of available data and meta-data, I did not apply any conversion factors to nutrient values for preparation method or specific gravity. While these methodological choices were consistent with those of my peers (Jenneson et al., 2021), they are likely to have contributed an over-estimation of nutrient values for diluted products, and an under-estimation in dense products sold by volume, such as ice cream and yoghurt.

6.2.3 Categorisation approaches

Dietary monitoring at the nutrient-level is important for assessing dietary adequacy and deficiency risk, but as described in the previous section I observed barriers to data linkage and accuracy. Much of the work in the field to date has instead reported dietary outcomes at the food category-level (Chapter 3 (Jenneson et al., 2021)). Aside from it being comparatively easier (avoiding the need for nutrient data linkage), monitoring purchases by food category is logical, given that people buy foods, not nutrients. Despite ongoing efforts to make nutritional labels more transparent (DHSC, 2020), understanding and use in purchase decision-making is relatively low (Moore et al., 2018). Therefore, if we wish to understand dietary habits, and to influence dietary choices for the better, it is important to think in terms of food groups, not just nutrients. Indeed, there has been an increase globally in national Food-Based Dietary Guidelines (WHO, 2021), which recognises the ease of category-level dietary communication and monitoring.

With in-built food categories, purchase records provide potential metrics for dietary patterns (Clark, et al., 2021), dietary quality (Appelhans et al., 2017) and for assessment against Food-Based Dietary Guidelines (Clark, et al., 2020). However, retailer categories are not nutritionally aligned. Instead, categories are led by store placement, business structure, or product use, which results in nutritional heterogeneity (e.g. ‘frozen foods’ includes vegetables, ready meals and ice cream etc.), and the dispersion of food groups of interest across several retail categories (for example, vegetables may be found in the ‘fresh produce’, ‘frozen foods’ and ‘tinned foods’ categories). Furthermore, categorisation approaches differ between retailers and are subject to change as new products emerge onto the market, such as we have seen with the growing availability of plant-based meat and dairy alternatives.

This lack of alignment causes problems for accuracy and between-study comparison. To resolve this, the retail data for my analysis was mapped to a new set of categories based on the Living Costs and Food Survey (Office for National Statistics., 2017), which allowed for identification of all fruits and vegetables sold fresh, frozen or canned, and the exclusion of potatoes, in line with UK 5-a-day portion recommendations (NHS, 2018). However, it did not capture fruits and vegetables in composite dishes (Jenneson, et al., 2022), resulting in an under-estimation of purchased fruit and vegetables. While Sainsbury's estimates the vegetable content of composite dishes to contribute relatively little to overall vegetable intake (unpublished data), under-estimation is likely to be biased among busy families and low-income groups who engage less frequently in scratch-cooking (Winkler and Turrell, 2010; Mills, 2018; Adams et al., 2015). Issues of categorisation and disaggregation are therefore important as they have implications for equity of dietary representation and could translate to spatial differences in representativeness of the metric.

The exact composition of products is proprietary information and cannot be shared. While methods to estimate product composition from the ingredients list and nutrition information (Bandy et al., 2021) are useful in research, a lack of transparency still poses a challenge for legislative compliance (Jenneson et al., 2020; Jenneson and Morris, 2021). Upcoming promotional restrictions by location and price (DHSC, 2019), will necessitate estimation of the fruit, vegetable and nut content of products to accurately calculate their UK Nutrient Profiling Model (NPM) score (Jenneson et al., 2020; DH, 2011). Furthermore, recommendation 12 of the National Food Strategy report calls for industry-wide reporting of food sales against a range of nutritional metrics (National Food Strategy., 2021). Therefore, while a gap currently exists for the use of retail transaction data to support policy, it is possible that the need to align to regulative standards and reporting may improve the availability and transparency of data. The mechanism and infrastructure for sharing such data are open for debate. However, options may include: the incorporation of additional data fields into product barcode standards (GS1, 2021b); and industry-wide incorporation of a standardised categorisation approach. Potential categorisation systems already in existence include; those which meet business needs e.g. GS1's Global Product Classification (GPC) (GS1, 2021a) and LANGUAL (2017); regulatory categories such as FoodEx2 (EFSA, 2015); and those used by government statistical agencies such as COICOP (DESA, 2018), CoFID (PHE, 2021) and others. Practical developments in this area would benefit not only researchers, but policy-makers and industry actors too.

6.3 Spatial granularity

The FED, introduced in Chapter 1 of this thesis, sets out the importance of the food environment as an up-stream influencer of food purchases and subsequent intake. Thus, as each place has its own unique food environment, one may consider the diets of individuals living there to be inherently place-based. Yet, exploration of the geographic nature of dietary behaviours has been hindered to date by the spatial granularity of dietary intake data. Small sample sizes limit geographic granularity to the regional level and require pooling of data over several years, reducing responsiveness to temporal trends. Exploration of food environments have revealed spatial patterns in outlet coverage and food availability at the neighbourhood level (Fraser et al., 2010; Blake, 2019) which suggest dietary exposures are inequitably distributed. Yet, the link between neighbourhood food environment exposures and actual dietary behaviours cannot readily be made at scale.

Despite their potential, my synthesis of the literature in Chapter 3, found spatial exploration of supermarket purchase records to be notably under-explored (Jenneson et al., 2021). Uniquely, my sample of loyalty card purchase records provides data for customer-level dietary behaviours, geocoded at both the store and customer-level. This offers a novel opportunity to study the interaction between diet, person, and place within a single dataset, contributing to knowledge of the small-area geographies of diet.

6.3.1 Contribution to ecological research

Ecological research is concerned with how exposures, health outcomes, and the interactions between them vary over space. It is important for hypothesis generation, identification of at-risk groups and the targeting of interventions and services. By linking transactions at the customer level, loyalty card purchase records form a large-scale convenience cohort of customers who may be followed up over time. Purchases may be linked with customer demographic information collected at loyalty card sign-up. While coverage is basic (for example it is unlikely to include sensitive information such as income and ethnicity) and its use limited by issues with data quality and access, the geocoded nature of transactions permit linkage with area level population data as a surrogate for detailed customer information.

I found few examples of purchase data in ecological research to date (Jenneson et al., 2021). However, correlations between purchases of different nutrients and prevalence of overweight and obesity, and Type 2 diabetes across the city of London observed by Aiello et al., (2020), adds weight to the validity of transaction records as an ecological dietary indicator. While the scope of my research prohibited linkage with population health or wider food environment data, I was able to observe marked variations in household fruit and vegetable purchases across a single city, which may translate to variation in health outcomes (Chapter 4 (Jenneson, et al., 2022)). Linkage with both customer-level and area-level demographic data enabled me to interrogate the clustering of low fruit and vegetable purchasing in areas with higher deprivation and urban density, but also revealed neighbourhoods which opposed the generally accepted association between deprivation and poor dietary quality. By examining nuance in purchase behaviours, customer characteristics and area-level characteristics in tandem, I was able to hypothesise factors which might mediate the diet-deprivation relationship in certain locations, such as education, ethnicity, and cultural preferences.

Although it is not possible to conclude whether this spatial variation is simply a consequence of geographic differences in purchase coverage (Rains and Longley, 2021), my observations invite further interrogation. Indeed, differences in agreement between purchases and intake by household size and customer loyalty observed in the STRIDE study (Chapter 5), suggest that agreement is likely to vary spatially too. Autocorrelation, which sees people with shared characteristics living in close proximity with one another under Tobler’s first law of geography (Tobler, 1970), suggests we can expect to see poorer agreement in areas populated with larger households and areas with a higher concentration of competitor stores, suggestive of lower customer loyalty. Spatial exploration of agreement could thus shed additional light on whether observed geographic purchase patterns represent differences in dietary intake, or differences in shopping habits (Rains and Longley, 2021) (e.g. buying fresh produce from the local greengrocer or growing your own).

While not yet explored, it is also feasible that agreement may vary by customer demographic characteristics and dietary purchase habits which could also contribute to spatial variation in agreement. For example, as younger people consume a higher proportion of restaurant and takeaway meals (Adams et al., 2015), this may point to poorer agreement in areas with a younger age demographic. Practically, this would mean that purchase data is a better dietary indicator in some areas than others, the implications of which for local-level policy making therefore warrant further exploration. While data collected in the STRIDE study may contribute to this, its small sample size would limit spatial granularity. Exploration by geodemographic group (segmentation of areas according to aggregated demographic attributes) using the Output Area

Classification is thus more feasible. Together with the work by Aiello et al., (2020), my work provides a foundation to further explore the important contribution that purchase data, in combination with other data sources, can make to ecological study and the development and monitoring of place-based interventions to improve diet-related inequalities.

6.4 Population coverage

While the majority of people shop at supermarkets, different retailers tend to attract different customers due to their price-point, offering and values. As a result, it is unlikely that transactions for a given supermarket are representative of the overall population. Furthermore, the inclusion of loyalty card holders only, introduces further selection bias. Despite recognition of such biases, my literature review (Chapter 3 (Jenneson et al., 2021)) found that purchase data samples were, on the whole, poorly characterised raising questions around the generalisability of findings. What is apparent however, is that a pervasive culture of supermarket shopping is important within a population for findings from supermarket purchases to be meaningful. Thus, the applicability of supermarket purchase records as a dietary metric is unlikely to extend beyond high- and middle-income countries.

To address questions of generalisability, it is important to characterise the customer sample (Rains and Longley, 2021). The availability of customer demographic data may prohibit detailed understanding of customer samples (for example, income and educational attainment level), such is possible in survey samples, but area-level characteristics can offer a useful proxy. The Sainsbury’s loyalty card customer sample is described in Chapter 4 (Jenneson, et al., 2022) as being predominantly female, of a slightly older age demographic, and as living in more affluent areas (Clark, et al., 2021). Despite the apparent bias in customer demographics, an advantage of transaction data is their scale. As a result, even those under-represented customer groups (younger customers and those living in deprived and more ethnically diverse areas) were found to be present in large enough numbers (Clark, et al., 2021; Jenneson et al., 2022) to permit between-group comparisons and weighting to align with general population characteristics. This is an advantage over survey methods, where the recruitment of hard-to-reach groups has become a specialist sub-discipline and statistical power may be so limited as to prohibit subgroup analyses.

The ability to combine dietary insight from multiple retailers is desirable to maximise population coverage. Yet the feasibility of this is dependent upon the setting. Due to the retail duopoly in Finland, the LoCard study represents very good coverage of the population, with a 46% market share (Vepsalainen et al., 2021; Nevalainen et al., 2018; Uusitalo et al., 2019). But the UK market is much more dispersed. Sainsbury's is considered one of the big-four supermarket chains and has a market share of 15.3% (Statista, 2021), so a smaller proportion of the population and their shopping habits will be represented in their data. That said, the 2016 customer sample for Leeds represented upwards of 6% of the city's population ((Jenneson, et al., 2022) Chapter 4) (a conservative estimate assuming all customers are from single-person households), a reach much greater than national dietary surveys.

6.5 Representativeness of diet

According to the FFS, on average 78% of UK food expenditure is spent in large supermarkets (Office for National Statistics., 2021), making them an important contributor to the nation's diet. While supermarket transactions therefore have the potential to capture most of what we consume, an inherent problem is that they measure purchases at the household-level and only for a single retailer, thus perfect agreement between household purchases and individual intake cannot be expected. First introduced in Chapter 1 (Figure 1.1), the FED is reproduced here (Figure 6.1) to demonstrate this. Here, the yellow outline depicts an indicative flow of food from a single retailer, illustrating that it captures just a proportion of the food at each stage. Agreement is thus dependent upon: 1) the proportion of total purchases captured by retailer purchases (S1 Food Environment → S1 Food Purchase transition); 2) the contribution of other food sources to household availability (S1 Food Environment → S3 Household Food Availability transition); 3) the proportion of food which is wasted by the household (S3 Household Food Availability → S4 Food Consumption transition); 4) the proportion of food consumed by other household members and guests (S3 Household Food Availability → S4 Food Consumption transition); and 6) measurement error associated with the chosen dietary intake assessment method (bias in the estimation of S4 Food Consumption). The flow of foods and nutrients between each stage of the FED will vary by person and setting.

Despite the rise in applications of supermarket transactions data to dietary research, my appraisal of the literature found relatively few validation studies to date. Just one study included in my systematic review directly compared electronically captured supermarket purchases and intake (Eyles et al., 2010), while another was published since the review was conducted (Vepsäläinen et al., 2021). Although both conclude that household purchases agree moderately well with individual intake (Vepsäläinen et al., 2021; Eyles et al., 2010), they are unable to indicate whether one method has a tendency to produce higher estimates than the other, and if so, by what magnitude. To address the need for a better understanding of the extent to which purchases represent intake, I conducted the STRIDE study. Using the Bland-Altman method (Bland, 1986; Bland, 1999), which is commonly regarded as the gold-standard method for comparing measures in health research and clinical practice, the STRIDE study adds to current knowledge by quantifying agreement and limits of agreement between the methods. For the first time, the STRIDE study compares absolute quantities of purchased and consumed nutrients, both at the household and the individual level. In the next sections I explain how, with the help of additional survey data, the STRIDE study aimed to account for some of the transitions through the FED which add up to discrepancies between purchase and intake. The methods and findings are discussed in relation to the wider literature and applicability to policy.

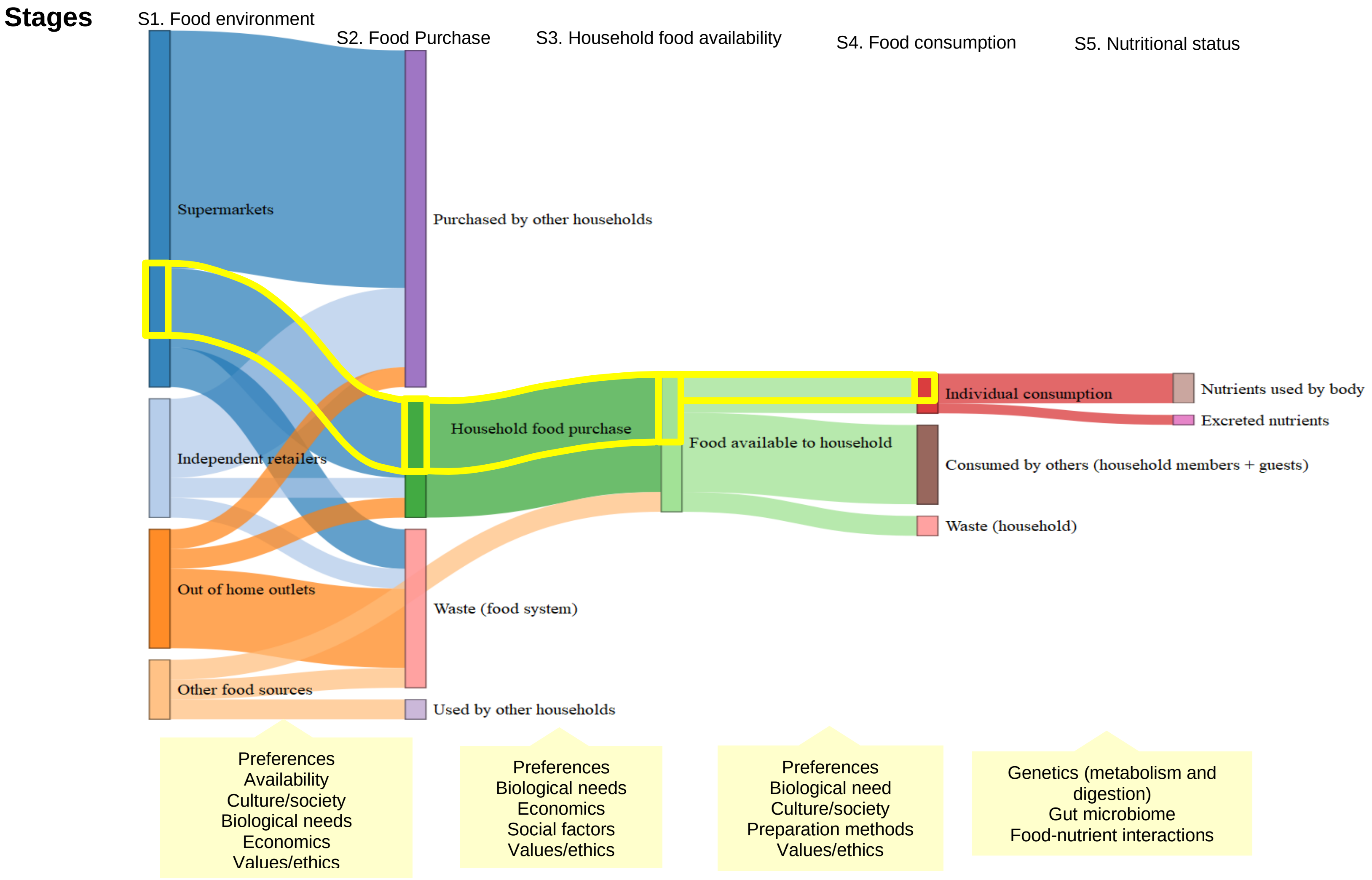


Figure 6.1 Framework for Evaluating Diet (FED) showing indicative proportion of food from a single retailer, outlined in yellow.

6.5.1 Coverage of total purchases

At the individual-level, Ransley et al. (2001) found total food purchases (measured by paper receipts collected from all sources including independent retailers and the out of home sector) to agree very well with intake recorded by food diaries (S2 → S4 transition in the FED). Covering just a proportion of total purchases (FED Stage 2), household purchase records from Sainsbury's collected by the STRIDE study captured on average 66% of the energy consumed by study participants. In other words, household purchases showed a tendency to under-estimate individual energy intake by around a third. Had this been a comparison between individual-level purchases and intake, this difference would be explained by the contribution of other food sources such as smaller retailers, home-grown produce, and the out of home sector, which were not captured by the STRIDE study. Yet, given that many households contain more than one person, it is surprising to see average household purchases under-estimating individual intake. Indeed, the variability in agreement by magnitude of the mean of methods indicates that mean agreement is meaningless unless presented as a regression equation which accounts for magnitude. Agreement with intake was found to be highly variable at the customer-level and thus household purchases cannot be considered a good proxy for intake for individual customers.

Household purchase estimates are a more reliable indicator of intake for smaller households, as purchased food is theoretically more likely to be consumed by the study participant. Thus, one might expect to find better agreement for single-person households than the population as a whole. However, this was not the case. A number of STRIDE participants were found to have extremely low daily energy purchase estimates, indicating low loyalty to the retailer and an under-estimation of intake. Other customers had high energy purchases, which may lead to over-estimation of intake due to large household sizes, high volumes of food waste, or purchasing for other households (a phenomenon which saw an increase during the COVID-19 pandemic as people supported others who were isolating (PHE, 2020a), indeed 13% of STRIDE respondents reported such a change).

Findings from STRIDE concur with the FED model. Agreement with intake is dependent on, the proportion of purchases captured by the retailer (S1 → S2 FED transition) (thus I observe variation by customer loyalty and purchase magnitude) and the amount of available food consumed by the individual (S3→ S4 FED transition). Thus, agreement varies by household size and magnitude. Whilst not explored directly by the STRIDE study, it is also likely that agreement will vary at the customer-level by household composition (Vepsalainen et al., 2021; Eyles et al., 2010) and customer characteristics (such as age, gender, income and BMI), and thus warrants further exploration.

As demonstrated by the sensitivity analysis performed in the STRIDE study (which removed customers with low mean daily energy purchases), the reliability of absolute dietary estimates from purchase records may be improved by selecting loyal customer samples for whom purchases with the retailer capture more of their total food purchases (S2). Rains and Longley (2021) propose a method for assessing completeness at the customer-level by comparing purchases with average household-level purchases in the LCFS (Office for National Statistics., 2017), accounting for both breadth and volume of purchases across different categories. They argue that completeness also depends on store network coverage and consider Sainsbury's purchases to offer a 'complete' representation of purchased food for just 11.5% of loyalty card holders across the UK (61% of customers were considered 'incomplete'). While my 2016 sample (Chapter 4) was also selected based on breadth and frequency of purchases, customer completeness was not assessed against national survey data, nor was any other metric of loyalty (e.g. Recency, Frequency, or Monetary value (known as the RFM index)). Thus, it is unclear to what degree Sainsbury's purchases capture their total shopping.

Sampling could thus be key to the utility of supermarket purchase records in population dietary monitoring. Selecting 'complete' loyalty card customers for inclusion in analyses would give greater confidence in dietary estimates. Of course, such a sample would still require characterisation to assess generalisability. The remaining sample of complete customers predicted by Rains and Longley (2021) (1.2 - 1.3 million nationwide) would still be large enough to offer detailed customer and spatial insight, with the capacity to generate synthetic estimates for areas with low coverage. The STRIDE study provides evidence that, whichever criterion is used to select 'complete' customers, continued reassessment against it is required. Despite selecting loyal customers for recruitment, for whom it was deemed purchases would be somewhat representative of their overall diet, based on their baseline purchase behaviours (2019 calendar year), a number of customers recorded very low purchases in the study period (2020 and 2021) and would unlikely have met the criteria had it been applied prospectively. While this may well have been exacerbated by the COVID-19 pandemic (the high demand for online retail may have resulted in customers signing up to retailers that they didn’t usually visit in-store), it highlights how much customer purchase behaviours and retailer loyalty may change over time.

6.5.2 Extrapolation to the individual-level

In the absence of detailed consumption data for all household members, as collected by Ransley et al., (2001), previous comparison studies have not attempted to extrapolate purchases to the individual-level for comparison with intake. Instead, they have used relative measures which describe the composition of the diet, including energy-adjusted nutrient density (Eyles et al., 2010; Appelhans et al., 2017), relative purchase volumes (Vepsalainen et al., 2021), and dietary quality (Appelhans et al., 2017). Such measures allow for direct comparison between the household and individual-level regardless of absolute volumes. For the first time, the STRIDE study assesses agreement between absolute values for nutrients purchased and consumed at the individual-level, by 1) performing sub-group analysis for single-person households, and 2) extrapolating purchases to the individual-level based on household composition and energy requirements data. This goes some way to accounting for the transition between S3 (household food availability) and S4 (individual intake) in the FED, with the exception of consumption by guests and food waste. Although without FFQ data for other household members, estimates are somewhat crude, limited by a lack of gender information for other household members and the use of equal allocation of nutrients regardless of food source (for example, energy from purchased alcohol will be allocated to both children and adults, although it is most likely consumed by just the adults).

As expected, error in individual-purchase estimates increased with household size, further under-estimating energy intake. For single-person households as the food is not shared by other household members, the remaining energy is most likely from other food sources (e.g. from other retailers or the out of home sector). Assuming all households receive a similar proportion of their food from the retailer, any additional error must be due to noise from food distribution within the household and waste. It should also be noted that household size may not necessarily reflect the number of people sharing the purchased food. Large households may represent houses of multiple occupancy, such as student housing and many professional letting agreements in cities. In such cases, despite living in a large household, residents are likely to act like single-person households, buying just enough for themselves. Extrapolation to the individual-level would therefore lead to under-estimation compared with intake for these customers.

While the method for allocating nutrients may be refined, it will always contribute error to individual-level purchase estimates as we cannot know exactly how foods are distributed between household members. Furthermore, it should be acknowledged that purchase records are a secondary data source and do not contain information about household composition inherently. Collecting such information in a continued manner would require huge investment in surveys, which would limit the large sample sizes and cost effectiveness that make purchase data so attractive as a dietary assessment tool. To ensure future utility of the method, efforts are therefore needed to explore how household composition may be estimated in the absence of survey data. Microsimulation techniques (Robards et al., 2017) may be used to model household composition using census data. Alternatively, data-led approaches could be used to estimate household size and composition according to the patterns, volumes, and products purchased. For example, the presence of baby food or infant formula in the transaction data would indicate the customer is part of a young family household. The approach may be further strengthened by the inclusion of non-food indicators such as sanitary products, which would indicate the presence of a woman of childbearing age in the household. While this technique is commonly used by retailers to segment customers for targeted marketing (Harris et al., 2005), due to its proprietary nature it could not be shared. Furthermore, the use of additional data should be balanced with ethical considerations for the need and application.

6.5.3 Bias in consumption estimation

As no dietary estimate is without error, disagreement between purchase and intake will in part be due to bias in the dietary intake assessment method. The choice of dietary intake measurement tool was constrained for the STRIDE study by the need to select one which was both scientifically validated, and complied with industry standards for data security. Validation of the SCG FFQ shows it has a tendency to over-estimate energy and nutrient intake in relation to food diaries (Hollis et al., 2017; Mohd-Shukri et al., 2013), particularly among pregnant women with obesity (Mohd-Shukri et al., 2013). It is therefore possible that greater agreement may have been seen had another dietary intake assessment method be used. For example, closer agreement may be expected with the myfood24 tool (Carter et al., 2015) due to increased accuracy of intake at the product-level. However, as FFQs are considered the best indicator of habitual dietary intake (Shim et al., 2014), they may show better agreement for longer transaction periods, as observed by Vepsalainen et al., (2021). That said, the moderate agreement found by the STRIDE study is in line with previous validation studies which compared purchases against repeated 24-hour recalls (Eyles et al., 2010) and an alternative FFQ tool (Vepsalainen et al., 2021).

A limitation of the SCG FFQ tool was in its user-friendliness, which contributed to the lower-than-expected completion rate. While under half of participants who signed up to the study completed the FFQ, the proportion who started it was higher, with a number of people completing only the first page. In addition, difficulties completing the FFQ was a common reason given for withdrawal from the study among the small number of participants who actively withdrew. It is possible that this observation is particular to the older age demographic of study participants, as studies have found that older adults struggle more with the completion of online tools (Carter et al., 2015). That said, the mean age of participants did not change much from sign-up to completion. Future validation studies should consider comparing the same loyalty card transaction data source against different dietary intake methods, and with nutritional biomarkers.

6.5.4 Agreement by dietary metric

The STRIDE study found closer agreement with intake for relative measures than for absolute volumes of purchased nutrients. With the strongest agreement for fat and sugar,, agreement for relative nutrient values observed by the STRIDE study is in line with the magnitudes of agreement reported by Appelhans et al., (2017) and Ransley et al., (2001). While relative dietary measures are a useful indicator of dietary quality, absolute measures are needed to offer insight into dietary adequacy. That is, whether individuals are likely to be experiencing under- or over-nutrition in a given food group or nutrient. The STRIDE study is unique in its approach to extrapolate household purchased nutrient quantities to the individual-level, allowing comparison of nutrient volume. While absolute purchase estimates appear to be a comparatively poor marker of intake, amends to the method for individual-level extrapolation may deliver improvements.

The STRIDE study found agreement between purchase data and intake to be stronger for energy, total fat and sugar (Chapter 5). This concurs in part with findings by Eyles et al., (2010) who found household supermarket purchases and intake to correlate more strongly for the proportion of energy from total fat and saturated fat, compared with other nutrients. Conversely, another study comparing intake with household purchases from all sources (not just from a single retailer) reported lower correlations for nutrient densities for fat, saturated fat and sodium, compared with carbohydrates, protein, sugar, fibre, whole fruits, and vegetables (Appelhans et al., 2017). This may reflect the sources of high-fat staple products such as milk, cheese, cooking oil, tend to be purchased from supermarkets. Despite poorer agreement for other macronutrients reported by the STRIDE study, median values for the two methods are within a couple of percentage points of one another, consistent with findings by Eyles et al., (2010) and Green et al., (2020). Moderate correlation was also found at the food-category level (Vepsalainen et al., 2021; Appelhans et al., 2017), though quantification of agreement is warranted. Together, these findings suggest that household purchase records provide a good proxy for dietary composition and support their use in studies of dietary quality (Appelhans et al., 2017), dietary patterns (Clark, et al., 2021), and compliance with food-based dietary guidelines (Clark, et al., 2020).

Findings from the STRIDE study, together with those from Vepsalainen et al., (2021), suggest that transaction records agree better for staple food items than for bulk or store-cupboard items such as cooking oil and salt. Given different items are purchased with different frequencies, large volumes of purchase data are likely to give a better account of habitual diet. Indeed, Vepsalainen et al., (2021) reported greater agreement with intake for 12 months of purchase records compared with just one month. Further study is thus required to define the optimal amount of time required for purchase coverage to represent habitual intake and whether there is seasonal variation in agreement. Indeed, purchase patterns observed in loyalty card transactions may offer an alternative view on what constitutes ‘habitual’ diet, which has been somewhat constrained by our feasibility to measure individual intake over time.

6.6 The importance of partnerships

The food environment is increasingly becoming a focus of policy attention for population and planetary health. Indeed, the National Food Strategy., (2021), a recent independent review of the UK food system, advocates that further policy levers be applied to ‘break the junk food cycle’ which dominates our current food environment. In the UK, policy objectives have thus far been directed at: 1) making less-healthy foods more expensive (for example, the Soft Drinks Industry Levy (SDIL) (HMRC, 2018)); 2) making the product offering healthier (such as voluntary reformulation targets) (PHE, 2018; DOH, 2015); 3) making less-healthy products less prominent (television advertising watershed (ASA, 2018) and upcoming HFSS legislation for product placement and promotions (DHSC, 2019)); and 4) making product nutritional information more transparent for customers (mandatory back of pack (Department of Health., 2017) and voluntary front of pack labelling (which is currently under review (DHSC, 2020))). Policy activity highlights the need for effective monitoring to maximise learning for future practise, something which the Government has been criticised for neglecting in the past (Theis and White, 2021).

The desire for food system digitalisation has been brought into sharper relief by the system shocks felt due to the COVID-19 pandemic. As a result, the Feed-UK initiative has been proposed to develop a digital twin of the UK food system (Baty, 2020) which can be used in scenario modelling and hypothesis generation. Supermarket transaction records have an important place in such models, given the contribution which supermarkets make to the nation's diet. My work demonstrates the utility of purchase records as a marker of population diets and sets out how they may be utilized for public good, through national and local dietary monitoring and policymaking. This thesis therefore makes an important contribution to the national food digitalisation agenda.

With data infrastructure highlighted by the WHO as a barrier to widescale use of purchase metrics in dietary monitoring (WHO, 2021), the importance of strong academic-retailer partnerships should not be under-estimated. With funding from the Medical Research Council (MRC) and the Economic and Social Research Council (ESRC), the Leeds Institute for Data Analytics (LIDA) has developed a unique data infrastructure for secure storage of large commercial datasets. This, in combination with its experts in commercial big data analytics via the Consumer Data Research Centre (CDRC), has enabled LIDA to establish a trusted formalised partnership with Sainsbury's Plc for access to purchase data for dietary research (LIDA, 2021), upon which I was able to build. To ensure continued benefits to society, such partnerships must be nurtured. Capacity-building in the nutrition field in required in particular (de la Hunty et al., 2021); bringing in multi-disciplinary influences from data science, geography, behavioural sciences, and other fields will ensure nutrition insight can benefit from digital innovations. Through such partnerships it is possible that we may see in the future an incorporation of transactions data within wider diet and health research (Morris et al., 2018) and a better alignment of data tools and infrastructure which benefit both the public and private sectors.

6.7 Impacts of the COVID-19 pandemic

It is important to note that the STRIDE study took place in the context of the COVID-19 global pandemic. National restrictions upon the opening of hospitality venues and guidance to stay at home in the UK resulted in changes to the nation's shopping and eating habits. It is possible that these restrictions may have had a number of impacts upon the STRIDE study and the future generalisability of findings. Firstly, the pandemic may have affected recruitment figures. Sign-up figures peaked in the second cohort, then tailed off to their lowest level in cohort 4. This may reflect the nation's mood around utilising additional freedoms in the summer months, and fatigue around health studies and data sharing. A small number of participants also stated their dietary habits (and/or Sainsbury's purchases) not being reflective of the norm as a reason for withdrawal from the study.

Changes in dietary habits may have affected the observed agreement with intake during the pandemic. On the one hand, it is possible that the closure of the hospitality sector, restrictions on travel abroad, reductions in food waste, and an increased reliance upon supermarkets for food provision, resulted in observed agreement being closer than usual. On the other hand, trends for shopping locally, purchasing for other households (where members were shielding due to their vulnerability, or isolating due to contact with the virus), and a growth in the use of online delivery services and recipe box services may have reduced agreement. The growth of online retail could have worked in either direction. With a large number of people who had not previously used grocery ecommerce platforms signing up to these services, high demand meant that people could not always shop with the same retailer they would usually visit in store. This may have resulted in some customers becoming more loyal to Sainsbury's, while others may have gone elsewhere. This may explain why a number of STRIDE participants, who were selected based on the frequency and breadth of their 2019 purchases, had very low levels of purchasing with Sainsbury's in the 2020-2021 study period. What is more, as lockdown restrictions and social distancing guidelines changed throughout the year, it may prove difficult to disentangle seasonal trends from the impacts of changing restrictions.

6.8 Future research directions

The work in this thesis sets a foundation for future research to enhance understanding of the utility of supermarket loyalty card transaction records as a population dietary measure and to realise their potential through discussion of the practical limitations. Here I give a summary of the future research priorities as I see them, which fall into either: 1) Methodological explorations (a set of immediate-term priorities to enhance understanding and utility for research); or 2) Applications for longitudinal population dietary surveillance (a set of longer-terms practical requirements to actualise their potential as a continuous dietary assessment tool for research and policy).

1) Methodological exploration:
    - **Validation**: quantification of agreement by food group; assessment of agreement by customer characteristics (e.g. age, gender, ethnicity, dietary patterns etc); explorations of spatial patterns of agreement; validation against other dietary assessment methods, including biomarkers.
    - **Characterising households**: data-driven methods to understand household size and composition (without the need for surveys); methods to improve extrapolation of purchases to the household level (i.e. allocating foods/nutrients between household members).
    - **Habitual diets**: understanding the required data volumes to capture usual diet, particularly accounting for store-cupboard items; exploring seasonal variation in agreement.

2) Applications for longitudinal population dietary surveillance:
    - **Protocols and best practice**: development of data-infrastructure and protocols around data sharing, confidentiality and data ethics; linkage of commercial and open-source data; automation of processes for rapid insight and dissemination.
    - **Customer sampling**: developing sampling strategies to select a well-characterised national pool of 'complete' customers for longitudinal follow-up; incorporation of microsimulation techniques to better-represent population sub-groups and geographic areas with low coverage.
    - **Ecological applications**: area-level linkage with population health, food environment, and demographics data for ecological study; incorporation into food-system digitalisation infrastructure for modelling; linkage with individual-level health and lifestyle data.

6.9 Conclusions

Supermarket purchases contribute a significant amount to the nation's diet and thus provide a useful tool for population dietary surveillance, policy evaluation, ecological research and monitoring intervention success. Purchasing has an important place in the Framework for Evaluating Diet to understand more about how individuals make dietary choices within their food environment and to intervene for positive health and sustainability impacts. Loyalty cards provide a large national convenience sample of objective dietary purchase data, without the burden of data collection being placed on the individual. This enables longitudinal follow-up, small-area dietary exploration and sub-group analyses, which have the capacity to enhance our understanding and tackling of dietary inequalities, both at the national and local-levels. Household purchases are a good indicator of dietary nutrient composition at the individual-level, making them particularly valuable for assessing dietary patterns, dietary quality, and longitudinal population-level trends. By quantifying agreement, this work goes some way towards the generation of correction factors to allow translation between purchase and intake metrics. Purchases are better-indicative of intake among smaller households and loyal customers, and for energy, fat, and sugar. Further work is required to understand their limitations in other customer types and for other nutrients. Customer sampling is important for the reliability of dietary purchase estimates and thus efforts to select and characterise appropriate customer samples are required.

References

Adams, J., Goffe, L., Brown, T., Lake, A.A., Summerbell, C., White, M., Wrieden, W. and Adamson, A.J. 2015. Frequency and socio-demographic correlates of eating meals out and take-away meals at home: cross-sectional analysis of the UK national diet and nutrition survey, waves 1–4 (2008–12). *International Journal of Behavioral Nutrition and Physical Activity.* **12**(1), p51.

Aiello, L.M., Quercia, D., Schifanella, R. and Del Prete, L. 2020. Tesco Grocery 1.0, a large-scale dataset of grocery purchases in London. *Scientific Data.* **7**(1), p57.

An, R., Patel, D., Segal, D. and Sturm, R. 2013. Eating better for less: A national discount program for healthy food purchases in South Africa. *American Journal of Health Behavior.* **37**(1), pp.56-61.

An, R. and Sturm, R. 2017. A cash-back rebate program for healthy food purchases in South Africa: Selection and program effects in self-reported diet patterns. *American Journal of Health Behavior.* **41**(2), pp.152-162.

Appelhans, B.M., French, S.A., Tangney, C.C., Powell, L.M. and Wang, Y. 2017. To what extent do food purchases reflect shoppers’ diet quality and nutrient intake? *International Journal of Behavioral Nutrition and Physical Activity.* **14**(1), p46.

ASA. 2018. *Food advertising: evidence-based rules for children's multimedia lives.* [Online]. [Accessed 20/05/20]. Available from: https://www.asa.org.uk/news/food-advertising-evidence-based-rules-for-children-s-multimedia-lives.html

Bandy, L.K., Hollowell, S., Harrington, R., Scarborough, P., Jebb, S. and Rayner, M. 2021. Assessing the healthiness of UK food companies' product portfolios using food sales and nutrient composition data. *PloS one.* **16**(8), pp.e0254833-e0254833.

Baty, S., Butchers, M., Moss, M. 2020. FeedUK – building resilience by digitising the food system. *Food Science and Technology.* **34**(4), pp.46-49.

Birkin, M., Wilkins, E. and Morris, M.A. 2019. Creating a long-term future for big data in obesity research. *International Journal of Obesity.* **43**(12), pp.2587-2592.

Blake, M.K. 2019. More than Just Food: Food Insecurity and Resilient Place Making through Community Self-Organising. *Sustainability.* **11**(10), p2942.

Bland, M.J., Altman, D.G. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* **i**, pp.307-310.

Bland, M.J., Altman, D.G. 1999. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research.* **8**, pp.135-160.

Carter, M., Albar, S., Morris, M., Mulla, U., Hancock, N., Evans, C., Alwan, N., Greenwood, D., Hardie, L., Frost, G., Wark, P. and Cade, J. 2015. Development of a UK Online 24-h Dietary Assessment Tool: myfood24. *Nutrients.* **7**(6), p4016.

Carter, M., Hancock, N., Albar, S., Brown, H., Greenwood, D., Hardie, L., Frost, G., Wark, P. and Cade, J. 2016. Development of a New Branded UK Food Composition Database for an Online Dietary Assessment Tool. *Nutrients.* **8**(8), p480.

Chidambaram, V., Brewster, P.J., Jordan, K.C. and Hurdle, J.F. 2013. qDIET: toward an automated, self-sustaining knowledge base to facilitate linking point-of-sale grocery items to nutritional content. *AMIA ... Annual Symposium proceedings. AMIA Symposium.* **2013**, pp.224-233.

Clark, S., Shute, B., Jenneson, V., Rains, T. and Morris, M. 2020. Compliance with the Eatwell guide: a case study using supermarket transaction records in Yorkshire and the Humber. *Proceedings of the Nutrition Society.* **79**(OCE2), pE665.

Clark, S.D., Shute, B., Jenneson, V., Rains, T., Birkin, M. and Morris, M.A. 2021. Dietary Patterns Derived from UK Supermarket Transaction Data with Nutrient and Socioeconomic Profiles. *Nutrients.* **13**(5), p1481.

de la Hunty, A., Buttriss, J., Draper, J., Roche, H., Levey, G., Florescu, A., Penfold, N. and Frost, G. 2021. UK Nutrition Research Partnership (NRP) workshop: Forum on advancing dietary intake assessment. *Nutrition Bulletin.* **46**(2), pp.228-237.

Department of Health. 2017. *Technical guidance on nutrition labelling.* London, UK: UK Government. [Online] [Accessed 13.06.2018]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/595961/Nutrition_Technical_Guidance.pdf

DESA. 2018. *Classification of Individual Consumption According to Purpose (COICOP) 2018*. Department of Economic and Social Affairs Statistics Division. New York: United Nations. Satistical Papers Available from: https://unstats.un.org/unsd/classifications/unsdclassifications/COICOP_2018_-_pre-edited_white_cover_version_-_2018-12-26.pdf

DH. 2011. *Nutrient Profiling Technical Guidance*. London: Crown copyright. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/216094/dh_123492.pdf

DHSC. 2019. *Consultation on restricting promotions of products high in fat, sugar and salt*. Department of Health and Social Care. London: Assets Publishing Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/770704/consultation-on-restricting-price-promotions-of-HFSS-products.pdf

DHSC. 2020. *Building on the success of front-of-pack nutrition labelling in the UK: a public consultation.* London.

DOH. 2015. *Public Health Responsibility Deal.* [Online]. [Accessed 13.06.2018]. Available from: http://webarchive.nationalarchives.gov.uk/20180201175643/https://responsibilitydeal.dh.gov.uk/

EFSA. 2015. *The food classification and description system FoodEx2 (revision 2)*. EFSA suppporting publication. European Food Safety Authority. pp.0-90. [Accessed 03.02.2022]. Available from: https://efsa.onlinelibrary.wiley.com/doi/epdf/10.2903/sp.efsa.2015.EN-804

Eyles, H., Jiang, Y. and Mhurchu, C.N. 2010. Use of Household Supermarket Sales Data to Estimate Nutrient Intakes: A Comparison with Repeat 24-Hour Dietary Recalls. *Journal of the American Dietetic Association.* **110**(1), pp.106-110.

Fong, M., Scott, S., Albani, V., Adamson, A. and Kaner, E. 2021. 'Joining the Dots': Individual, Sociocultural and Environmental Links between Alcohol Consumption, Dietary Intake and Body Weight—A Narrative Review. *Nutrients.* **13**(9), p2927.

Foster, J.H. and Ferguson, C.S. 2012. Home Drinking in the UK: Trends and Causes. *Alcohol and Alcoholism.* **47**(3), pp.355-358.

Fraser, L.K., Edwards, K.L., Cade, J. and Clarke, G.P. 2010. The Geography of Fast Food Outlets: A Review. *International Journal of Environmental Research and Public Health.* **7**(5), pp.2290-2308.

Green, M.A., Watson, A.W., Brunstrom, J.M., Corfe, B.M., Johnstone, A.M., Williams, E.A. and Stevenson, E. 2020. Comparing supermarket loyalty card data with traditional diet survey data for understanding how protein is purchased and consumed in older adults for the UK, 2014–16. *Nutrition Journal.* **19**(1), p83.

GS1. 2021a. *How Global Product Classification (GPC) works.* [Online]. [Accessed 02.03.2022]. Available from: https://www.gs1.org/standards/gpc/how-gpc-works

GS1. 2021b. *productDNA. How does productDNA work?* [Online]. [Accessed 02.03.2022]. Available from: https://productdna.gs1uk.org/how-it-works

Hamilton, S., Mhurchu, C.N. and Priest, P. 2007. Food and nutrient availability in New Zealand: An analysis of supermarket sales data. *Public Health Nutrition.* **10**(12), pp.1448-1455.

Harrington, R.A., Adhikari, V., Rayner, M. and Scarborough, P. 2019. Nutrient composition databases in the age of big data: foodDB, a comprehensive, real-time database infrastructure. *BMJ Open.* **9**(6), pe026652.

Harris, R., Sleight, P. and Webber, R. 2005. *Geodemographics, GIS and neighbourhood targeting.* John Wiley & Sons.

HMRC. 2018. *Check if your drink is liable for the Soft Drink Industry Levy.* [Online]. [Accessed 13.08.19]. Available from: https://www.gov.uk/guidance/check-if-your-drink-is-liable-for-the-soft-drinks-industry-levy

Hollis, J.L., Craig, L.C., Whybrow, S., Clark, H., Kyle, J.A. and McNeill, G. 2017. Assessing the relative validity of the Scottish Collaborative Group FFQ for measuring dietary intake in adults. *Public Health Nutr.* **20**(3), pp.449-455.

IAS. 2018. *Institute of Alcohol Studies Factsheet: The alcohol industry*. [Accessed 03.02.2022]. Available from:

https://www.ias.org.uk/uploads/pdf/Factsheets/FS%20industry%20012018.pdf

Institute for Government. 2021. *Timeline of UK coronavirus lockdowns, March 2020 to March 2021.* [Online]. [Accessed 24.01.2022]. Available from: https://www.instituteforgovernment.org.uk/sites/default/files/timeline-lockdown-web.pdf

Jenneson, V., Clarke, G.P., Greenwood, D.C., Shute, B., Tempest, B., Rains, T. and Morris, M.A. 2022. Exploring the Geographic Variation in Fruit and Vegetable Purchasing Behaviour Using Supermarket Transaction Data. *Nutrients.* **14**(1), p177.

Jenneson, V., Dyer, J., Matousek, A., Francois, I., Tumelty, J., Omieljaniuk, M., Stephens, M., Martin, R., Wu, S., Yung Low, S., Yip, W. and Morris, M.A. 2022 (forthcoming). *Investigating the impact of the UK's Soft Drinks Industry Levy on consumers' purchases of soft drinks.* London, UK: The Alan Turing Institute Data Study Group.

Jenneson, V., Greenwood, D.C., Clarke, G.P., Hancock, N., Cade, J.E. and Morris, M.A. 2020. Restricting promotions of ‘less healthy’ foods and beverages by price and location: A big data application of UK Nutrient Profiling Models to a retail product dataset. *Nutrition Bulletin.* **45**(4), pp.389-402.

Jenneson, V. and Morris, M.A. 2021. Data considerations for the success of policy to restrict in-store food promotions: A commentary from a food industry nutritionist consultation. *Nutrition Bulletin.* **46**(1), pp.40-51.

Jenneson, V., Pontin, F., Greenwood, D., Clarke, G. and Morris MA. 2021. A systematic review of automated electronic supermarket sales data for population dietary surveillance. *Nutrition Reviews.* nuab089

LANGUAL. 2017. *LANGUAL - The International Framework for Food*. Danish Food Informatics. [Accessed 03.02.2022]. Available from: https://www.langual.org/default.asp

LIDA. 2021. *LIDA announces partnership with Sainsbury's.* [Online]. [Accessed 24.01.2022]. Available from: https://lida.leeds.ac.uk/news/sainsburys-partnership/

Mills, S., Adams, J., Wrieden, W., White, M., Brown, H.,. 2018. Sociodemographic characteristics and frequency of consuming home-cooked meals and meals from out-of-home sources: cross-sectional analysis of a population-based cohort study. *Public Health Nutrition.*

Mohd-Shukri, N.A., Bolton, J.L., Norman, J.E., Walker, B.R. and Reynolds, R.M. 2013. Evaluation of an FFQ to assess total energy and nutrient intakes in severely obese pregnant women. *Public Health Nutrition.* **16**(8), pp.1427-1435.

Moore, S.G., Donnelly, J.K., Jones, S. and Cade, J.E. 2018. Effect of Educational Interventions on Understanding and Use of Nutrition Labels: A Systematic Review. *Nutrients.* **10**(10), p1432.

Morris, M., Glaser, A. and Iles-Smith, H. 2018. *Study protocol: a survey of public opinion on the acceptability of linking electronic health records with loyalty card (supermarket) and liefestyle app data for research.* LifeInfo

Survey: What do you think about researchers using your lifestyle information? : Open Science Framework.,.

National Food Strategy. 2021. *The Plan: National Food Strategy, Independent Review.* London, UK.

Nevalainen, J., Erkkola, M., Saarijärvi, H., Näppilä, T. and Fogelholm, M. 2018. Large-scale loyalty card data in health research. *Digital health.* **4**, pp.2055207618816898-2055207618816898.

NHS. 2018. *5 A Day: what counts?* [Online]. [Accessed 22/06/2020]. Available from: https://www.nhs.uk/live-well/eat-well/5-a-day-what-counts/

NHS. 2019. *The Eatwell Guide.* [Online]. [Accessed 03.02.2022]. Available from: https://www.nhs.uk/live-well/eat-well/the-eatwell-guide/

Office for National Statistics. 2017. *Living costs and food survey: user guidance and technical information for the Living Costs and Food Survey.* [Online]. [Accessed 12.01.2022]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/methodologies/livingcostsandfoodsurvey

Office for National Statistics. 2021. Family Spending: Workbook 1 - detailed expenditure and trends. Table A2. London, UK: *Office for National Statistics*. [Online]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/expenditure/datasets/familyspendingworkbook1detailedexpenditureandtrends

PHE. 2018. *Sugar reduction and wider reformulation programme: Report on progress towards the first 5% reduction and next steps*. PHE publications. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/709008/Sugar_reduction_progress_report.pdf

PHE. 2020a. *Impact of COVID-19 pandemic on grocery shopping behaviours.* London, UK.

PHE. 2020b. McCance and Widdowson's composition of foods integrated dataset. London, UK: *gov.uk*. [Online]. [Accessed 21.01.2022]. Available from: https://www.gov.uk/government/publications/composition-of-foods-integrated-dataset-cofid

PHE. 2021. *McCance and Widdowson's The Composition of Foods Integrated Dataset 2021*. User guide. UK: Assets Publishing. pp.23 - 26. Appendix B: Food sub-group codes. [Accessed 03.02.2022]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/971021/McCance_and_Widdowsons_Composition_of_Foods_integrated_dataset_2021.pdf

Piernas, C., Aveyard, P., Lee, C., Tsiountsioura, M., Noreik, M., Astbury, N.M., Oke, J., Madigan, C. and Jebb, S.A. 2020. Evaluation of an intervention to provide brief support and personalized feedback on food shopping to reduce saturated fat intake (PC-SHOP): A randomized controlled trial. *PLOS Medicine.* **17**(11), pe1003385.

Piernas, C., Tsiountsioura, M., Astbury, N.M., Madigan, C., Aveyard, P. and Jebb, S.A. 2019. Primary Care Shopping Intervention for Cardiovascular

Disease Prevention (PC-SHOP): protocol for a randomised controlled trial to reduce saturated fat intake. *BMJ Open.* **9**(4), pe027035.

Rains, T. and Longley, P. 2021. The provenance of loyalty card data for urban and retail analytics. *Journal of Retailing and Consumer Services.* **63**, p102650.

Ransley, J.K., Donnelly, J.K., Khara, T.N., Botham, H., Arnot, H., Greenwood, D.C. and Cade, J.E. 2001. The use of supermarket till receipts to determine the fat and energy intake in a UK population. *Public Health Nutrition.* **4**(6), pp.1279-1286.

Robards, J., Gale, C. and Martin, D. 2017. *Creating a synthetic spatial microdataset for zone design experiments using 2011 Census and linked administrative data.*

Shim, J.-S., Oh, K. and Kim, H.C. 2014. Dietary assessment methods in epidemiologic studies. *Epidemiology and health.* **36**, pp.e2014009-e2014009.

Statista. 2021. *Market share of grocery stores in Great Britain from January 2017 to May 2021.* [Online]. [Accessed 24.01.2022]. Available from: https://www.statista.com/statistics/280208/grocery-market-share-in-the-united-kingdom-uk/

Theis, D.R.Z. and White, M. 2021. Is Obesity Policy in England Fit for Purpose? Analysis of Government Strategies and Policies, 1992–2020. *The Milbank Quarterly.* **99**(1), pp.126-170.

Tobler, W.R. 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography.* **46**(2), pp.234-240.

Toft, U., Kristoffersen, L.H., Lau, C., Borch-Johnsen, K. and Jørgensen, T. 2007. The Dietary Quality Score: validation and association with cardiovascular risk factors: the Inter99 study. *European Journal of Clinical Nutrition.* **61**(2), pp.270-278.

Tran, L.T.T., Brewster, P.J., Chidambaram, V. and Hurdle, J.F. 2017. An innovative method for monitoring food quality and the healthfulness of consumers' grocery purchases. *Nutrients.* **9**(5), p457.

USDA. 2020. *Healthy Eating Index (HEI).* [Online]. [Accessed 03.02.2022]. Available from: https://www.fns.usda.gov/healthy-eating-index-hei

Uusitalo, L., Erkkola, M., Lintonen, T., Rahkonen, O. and Nevalainen, J. 2019. Alcohol expenditure in grocery stores and their associations with tobacco and food expenditures. *BMC Public Health.* **19**(787).

Vepsalainen, H., Nevalainen, J., Kinnunen, S., Itkonen, S.T., Meinila, J., Mannisto, S., Uusitalo, L., Fogelholm, M. and Erkkola, M. 2021. Do we eat what we buy? Relative validity of grocery purchase data as an indicator of food consumption in the LoCard study. *Br J Nutr.* pp.1-24.

Vepsäläinen, H., Nevalainen, J., Kinnunen, S., Itkonen, S.T., Meinilä, J., Männistö, S., Uusitalo, L., Fogelholm, M. and Erkkola, M. 2021. Do we eat what we buy? Relative validity of grocery purchase data as an indicator of food consumption in the LoCard study. *British Journal of Nutrition.* pp.1-24.

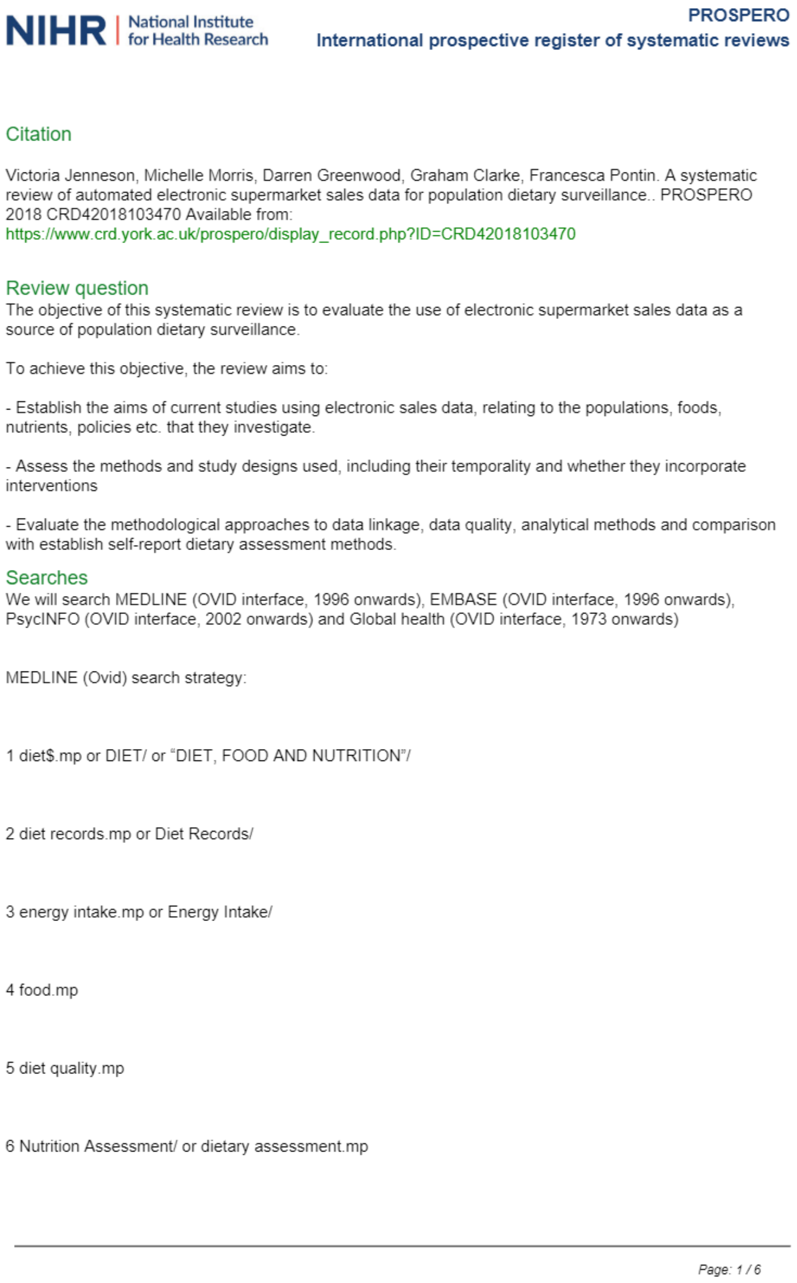
WHO. 2021. *Report of the technical consultation on measuring healthy diets: concepts, methods and metrics. Virtual meeting, 18 - 20 May 2021.* Geneva, Switzerland: World Health Organisation.

Winkler, E. and Turrell, G. 2010. Confidence to cook vegetables and the buying habits of Australian households. *J Am Diet Assoc.* **110**(5 Suppl), pp.S52-61.

Appendix A
Systematic Review supporting documents

Appendices in this section relate to paper 1, *‘A systematic review of automated electronic supermarket sales data for population dietary surveillance’*, presented in Chapter 3.

A.1 Systematic review protocol (as published on PROSPERO)



NIHR | National Institute for Health Research

PROSPERO
International prospective register of systematic reviews

Citation

Victoria Jenneson, Michelle Morris, Darren Greenwood, Graham Clarke, Francesca Pontin. A systematic review of automated electronic supermarket sales data for population dietary surveillance.. PROSPERO 2018 CRD42018103470 Available from:
https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42018103470

Review question

The objective of this systematic review is to evaluate the use of electronic supermarket sales data as a source of population dietary surveillance.

To achieve this objective, the review aims to:

- Establish the aims of current studies using electronic sales data, relating to the populations, foods, nutrients, policies etc. that they investigate.

- Assess the methods and study designs used, including their temporality and whether they incorporate interventions

- Evaluate the methodological approaches to data linkage, data quality, analytical methods and comparison with establish self-report dietary assessment methods.

Searches

We will search MEDLINE (OVID interface, 1996 onwards), EMBASE (OVID interface, 1996 onwards), PsycINFO (OVID interface, 2002 onwards) and Global health (OVID interface, 1973 onwards)

MEDLINE (Ovid) search strategy:

1 diet$.mp or DIET/ or “DIET, FOOD AND NUTRITION”/

2 diet records.mp or Diet Records/

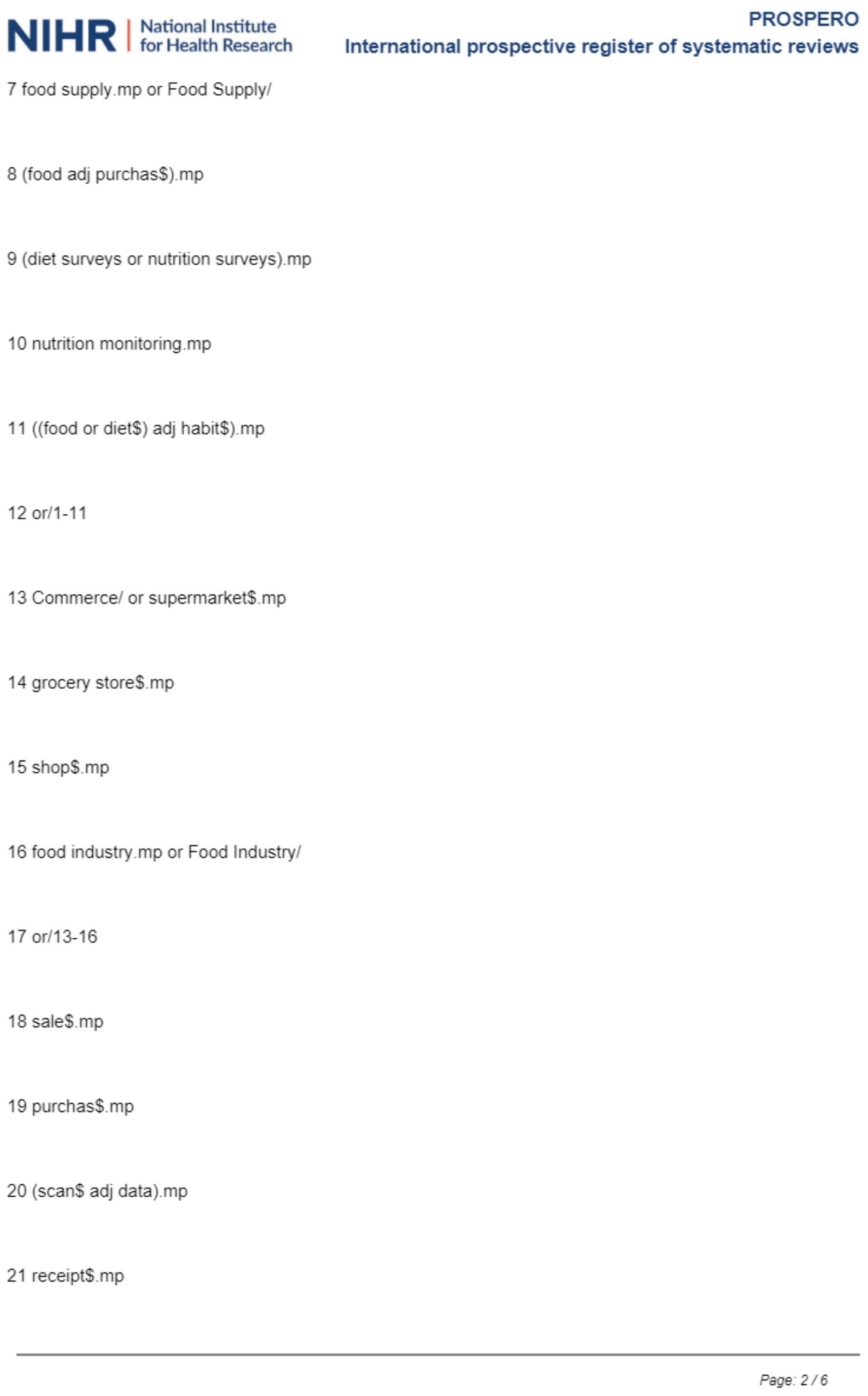
3 energy intake.mp or Energy Intake/

4 food.mp

5 diet quality.mp

6 Nutrition Assessment/ or dietary assessment.mp

Page: 1 / 6



7 food supply.mp or Food Supply/

8 (food adj purchas$).mp

9 (diet surveys or nutrition surveys).mp

10 nutrition monitoring.mp

11 ((food or diet$) adj habit$).mp

12 or/1-11

13 Commerce/ or supermarket$.mp

14 grocery store$.mp

15 shop$.mp

16 food industry.mp or Food Industry/

17 or/13-16

18 sale$.mp

19 purchas$.mp

20 (scan$ adj data).mp

21 receipt$.mp

22 (loyalty adj card).mp

23 or/18-22

24 and/12, 17, 23

Types of study to be included

No restrictions are placed on types of study design eligible for inclusion.

Condition or domain being studied

Food/nutritional purchases

Participants/population

Free-living individual adults or households (inclusive of those with children) will be included in the review. Studies exclusively on children, adults with known disease or pregnant women will be excluded.

Intervention(s), exposure(s)

Studies using electronic data captured as a direct result of food purchases, be they online or in store, will be included. Studies reliant on participant self-reports of food purchases will be excluded.

Comparator(s)/control

Not applicable

Context

In store or online electronic food purchase records.

Main outcome(s)

Quantity of food and beverage sales, at a product or category level (in monetary cost or volume terms)

Macro and micro-nutrient values purchased and/or consumed (e.g. kcal, grams as a total per day or per product/100g of product)

Measures of effect

Any

Additional outcome(s)

All other dietary-based outcomes which may include identification of dietary patterns and evaluation of public health nutrition policies.

Measures of effect

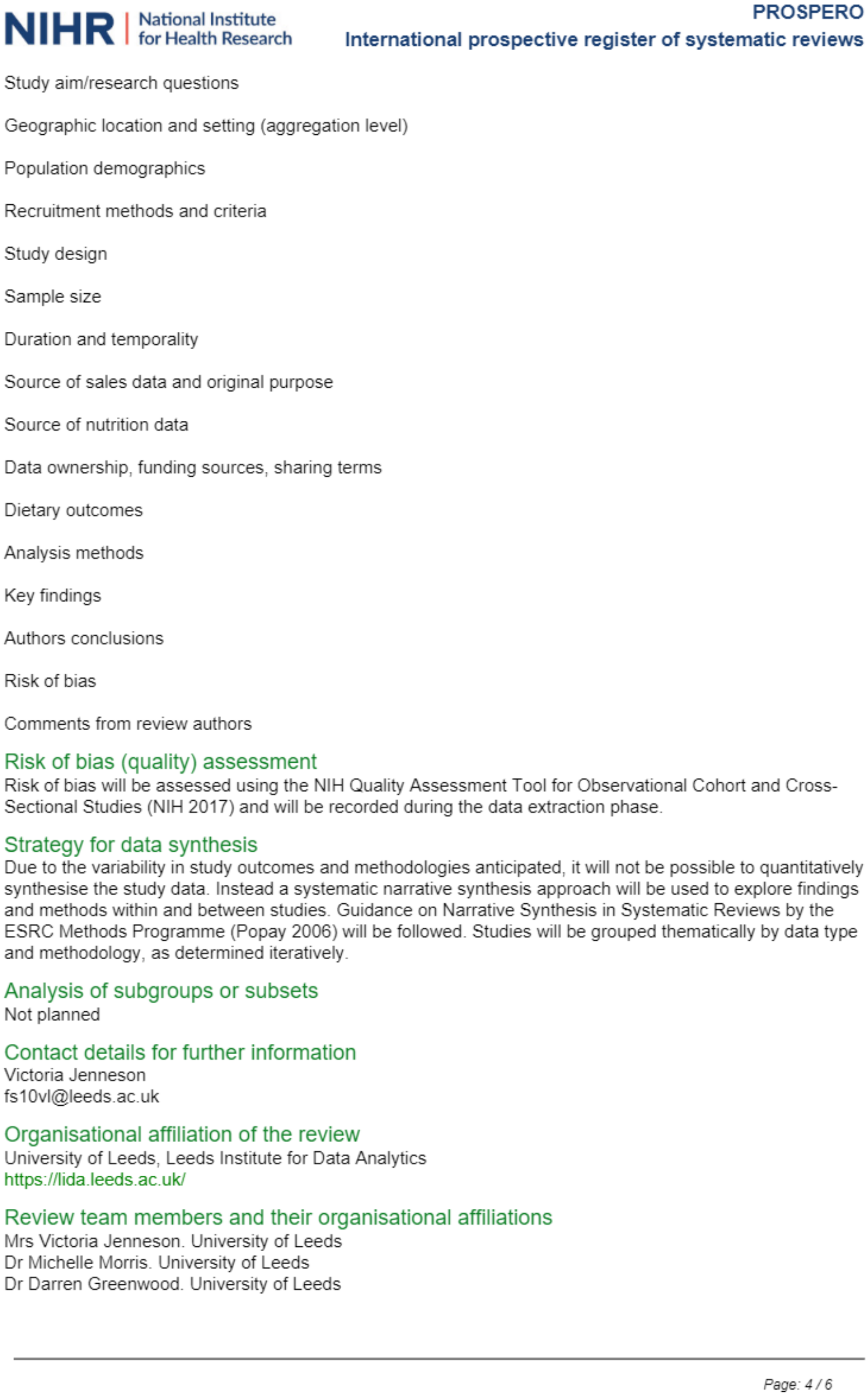
Any

Data extraction (selection and coding)

The following headings will be used for the pilot data extraction form, for testing. These data fields are adapted the Cochrane Public Health Group. (2011) data extraction form and the BEE COAST framework derived by Morris, Wilkins et al (unpublished work), which aims to characterise 'found' big data sources available in public health research, acknowledging their diversity.

Study ID (First author and date)

Title



Study aim/research questions

Geographic location and setting (aggregation level)

Population demographics

Recruitment methods and criteria

Study design

Sample size

Duration and temporality

Source of sales data and original purpose

Source of nutrition data

Data ownership, funding sources, sharing terms

Dietary outcomes

Analysis methods

Key findings

Authors conclusions

Risk of bias

Comments from review authors

Risk of bias (quality) assessment

Risk of bias will be assessed using the NIH Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies (NIH 2017) and will be recorded during the data extraction phase.

Strategy for data synthesis

Due to the variability in study outcomes and methodologies anticipated, it will not be possible to quantitatively synthesise the study data. Instead a systematic narrative synthesis approach will be used to explore findings and methods within and between studies. Guidance on Narrative Synthesis in Systematic Reviews by the ESRC Methods Programme (Popay 2006) will be followed. Studies will be grouped thematically by data type and methodology, as determined iteratively.

Analysis of subgroups or subsets

Not planned

Contact details for further information

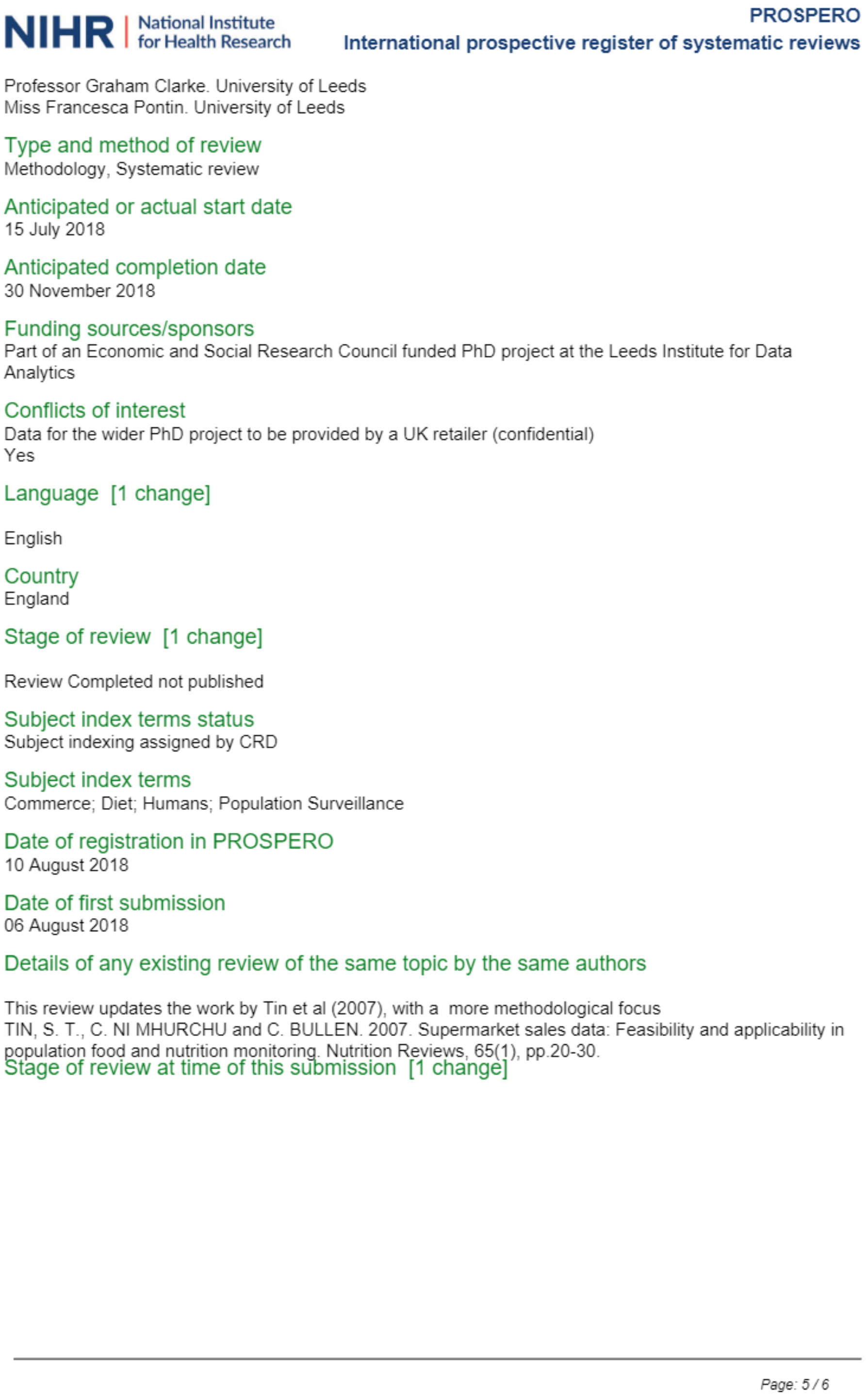
Victoria Jenneson
fs10vl@leeds.ac.uk

Organisational affiliation of the review

University of Leeds, Leeds Institute for Data Analytics
https://lida.leeds.ac.uk/

Review team members and their organisational affiliations

Mrs Victoria Jenneson. University of Leeds
Dr Michelle Morris. University of Leeds
Dr Darren Greenwood. University of Leeds



Professor Graham Clarke. University of Leeds
Miss Francesca Pontin. University of Leeds

Type and method of review
Methodology, Systematic review

Anticipated or actual start date
15 July 2018

Anticipated completion date
30 November 2018

Funding sources/sponsors
Part of an Economic and Social Research Council funded PhD project at the Leeds Institute for Data Analytics

Conflicts of interest
Data for the wider PhD project to be provided by a UK retailer (confidential)
Yes

Language [1 change]

English

Country
England

Stage of review [1 change]

Review Completed not published

Subject index terms status
Subject indexing assigned by CRD

Subject index terms
Commerce; Diet; Humans; Population Surveillance

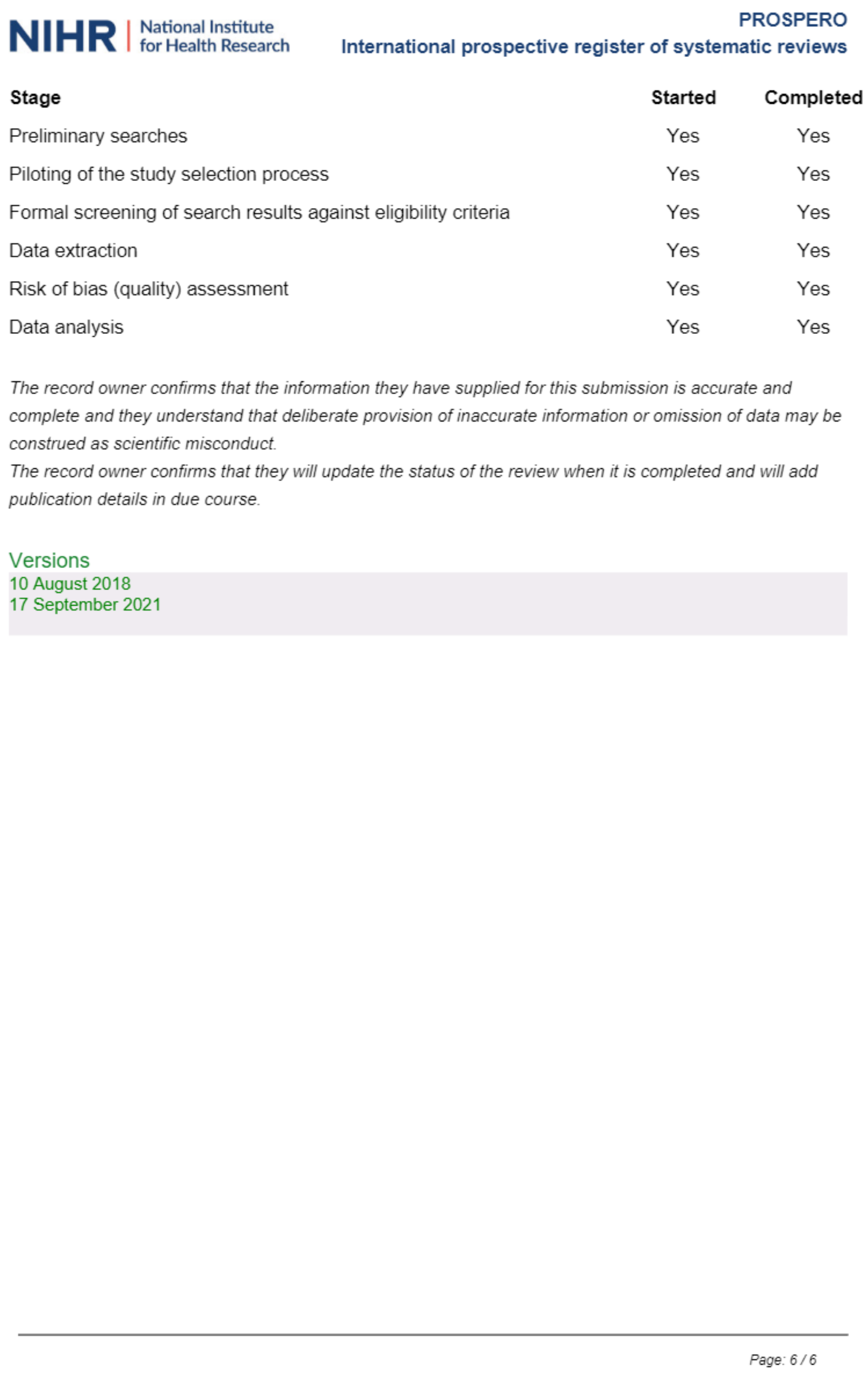
Date of registration in PROSPERO
10 August 2018

Date of first submission
06 August 2018

Details of any existing review of the same topic by the same authors

This review updates the work by Tin et al (2007), with a more methodological focus
TIN, S. T., C. NI MHURCHU and C. BULLEN. 2007. Supermarket sales data: Feasibility and applicability in population food and nutrition monitoring. Nutrition Reviews, 65(1), pp.20-30.

Stage of review at time of this submission [1 change]



Stage	Started	Completed
Preliminary searches	Yes	Yes
Piloting of the study selection process	Yes	Yes
Formal screening of search results against eligibility criteria	Yes	Yes
Data extraction	Yes	Yes
Risk of bias (quality) assessment	Yes	Yes
Data analysis	Yes	Yes

The record owner confirms that the information they have supplied for this submission is accurate and complete and they understand that deliberate provision of inaccurate information or omission of data may be construed as scientific misconduct.
The record owner confirms that they will update the status of the review when it is completed and will add publication details in due course.

Versions

10 August 2018
17 September 2021

Appendix B
STRIDE supporting documents

Appendices in this section relate to paper 3, *'Supermarket Transaction Records In Dietary Evaluation – The STRIDE study: validation against self-reported dietary intake*, presented in Chapter 5.

B.1 STRIDE data management plan

Project title and brief description (See note a.):	**STRIDE (Supermarket Transaction Records In Dietary Evaluation).** The aim of this project is to validate the use of electronically captured food purchase records as a source of population dietary information for nutrition research. The project will use secondary purchase data for a sample of Sainsbury's customers living in Yorkshire and the Humber. Sainsbury's will provide data on; participants' food and beverage transactions for the study period (September 2019 – November 2020) and for one year prior to participation in the study. Primary data will also be collected in digital format using an online survey provider and online dietary assessment tool. Data from the online tools will be downloaded into speadsheets within the Gateway, before entering the VRE. Primary and secondary data will enter the VRE as speadsheets (likely .xls or .csv file formats). R scripts and workspaces will be written within the VRE for analysis. Spatial analysis will also be conducted with QGIS and image files saved as outputs.
What data sets will you be using and from where? (See note e.)	The project will use a combination of primary data (collected via online survey platforms) and secondary data; transactions, customer information and product data provided by Sainsbury's. **Data to be provided by Sainsbury's:** • Unique ID which identified the nectar card and allows for linkage with nectar card ID, reported by participants. • Output Area for the customer home address. (Output Area is a unit of neighbourhood geography) • Demographic data for the Nectar card holder e.g. age range and gender. • Loyalty score for the Nectar card holder (where available)

	• Sainsbury’s store location and type (supermarket, convenience store or online) for each purchase. • Name of each item purchased. • Quantity of items purchased. • Cost of each item purchased. • Date and time of each item purchased. • Energy (Kcals), Fat, Saturates, Sugar, Salt, Protein and Fibre content of products purchased. • Pack weight All project participants will receive an initial contact email from Sainsbury’s containing a link to the study web page (hosted by the LIDA website) containing detailed participant information. The study web page will contain a link to an online survey platform for customers to complete an active informed consent form, giving their permission for their primary and secondary data to be used as outlined in the project summary. More details can be found in the Participant Information Sheet and STRIDE study Protocol attached. A data sharing agreement has been signed by Sainsbury’s and LIDA for this project. Sainsbury’s agree to share the secondary datasets requested and will undertake initial contact of prospective participants. Only upon completion of online consent forms will participant details be made available to authorised project researchers at LIDA. If a data breach should occur, the following parties should be informed; Victoria Jenneson, (lead investigator) Michelle Morris, (PhD supervisor) Becky Shute, (lead project contact at Sainsbury’s)
What files will be generated as part of the research? (See note c.)	5 cohorts of participants will take part in the study (Pilot + cohorts 1, 2, 3 and 4). Recruitment of the cohorts will be staggered over the year to ensure that dietary recalls reflect seasonal variation in diet. Around 400 people are expected to complete primary data collection for each cohort (200 for pilot), therefore data will be stored for around 1,800 people in total. This will include baseline survey data and informed consent, 4 x 24-hr dietary recalls, 2 years transaction data (study year + 1 year prior to study participation) and customer demographic data from the retailer for each participant. In addition, product nutrition data will be provided by the retailer.

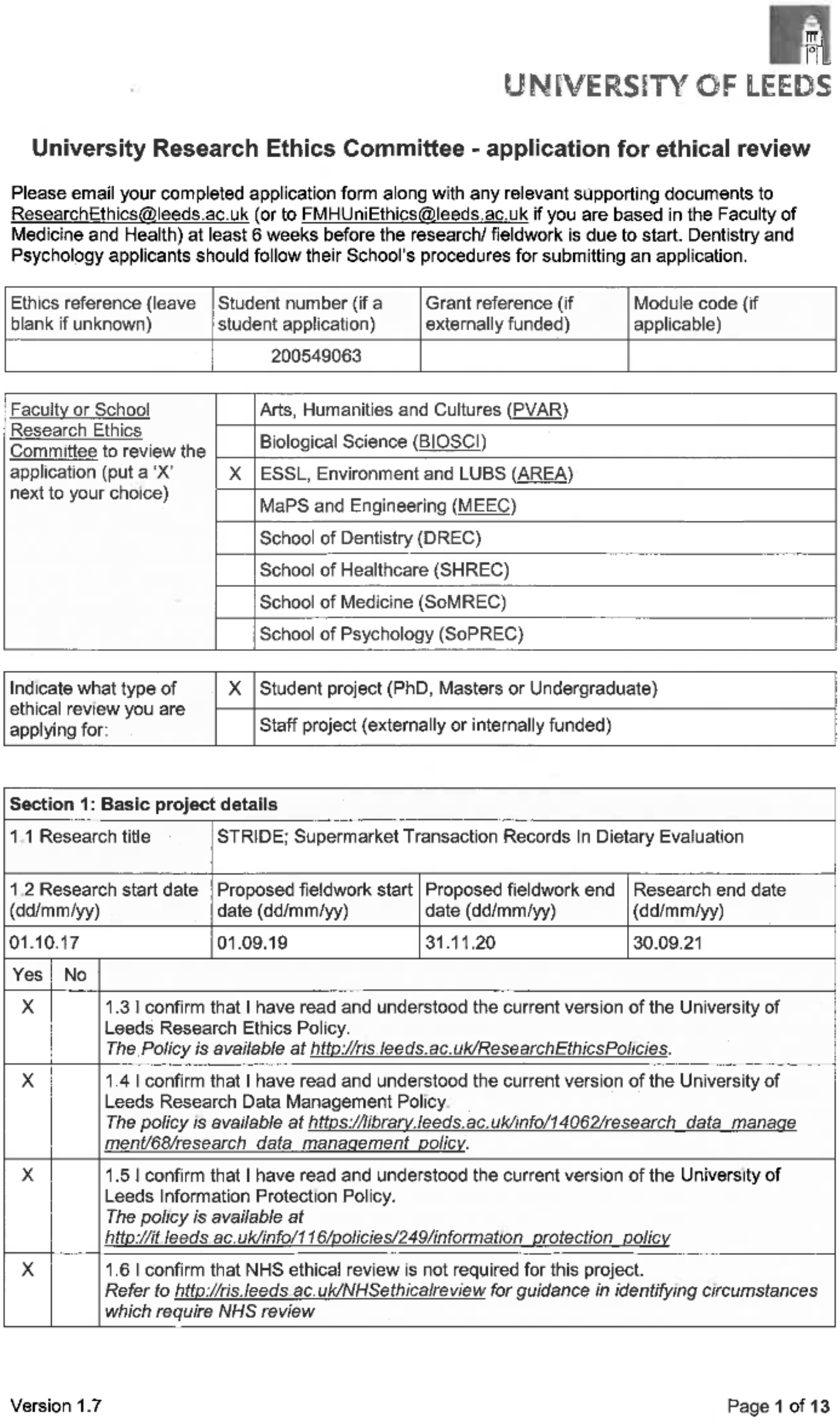
	It is intended that data from online survey platforms be accessed and downloaded directly into the Gateway, then transferred into the VRE to maximise data security. A list of participant email addresses will be required outside of the VRE in order for the researcher to send links to complete online surveys. A spreadsheet containing participant email addresses will be stored in a secure project folder on the University N-drive. Advice is being sought from central university IT about the level of security/encryption required for this data.
How will generated files be documented and described? (See note c.)	All participant data files will be named according to their contents and participation cohort. E.g. pilot_baseline_consent Files entering the VRE will be stored in dated incoming folders. File nomenclature will remain consistent across all cohorts and study phases. Meta data files will be generated within the VRE to describe data fields; this will only be required once as data field names will be consistent across all cohorts and study phases. Participants will be identifiable for linkage by their psuedonymised customer ID number. A data linkage key will be provided by Sainsbury's to link loyalty card numbers (provided by participants) with customer ID (provided by Sainsbury's)
How will your files be structured and stored? (See note c.)	Files will be stored in folders relating to the cohort and within each cohort folder, separate files for each study stage e.g. pilot → pilot_recall1
Are there any 'special' requirements for your information? (See note d.)	The project data will contain sensitive personal information including ethnicity, BMI and purchase behaviours. Primary data collected by researchers during the study, will not be shared with Sainsbury's to prevent its use for marketing purposes. Aggregated results from analyses will be shared with Sainsbury's. Sainsbury's will remove all purchase records relating to non-food items before sharing purchase data with the research team. This will reduce the risk of identification or revealing sensitive personal information e.g. whether a customer has recently taken a pregnancy test or follows particular religious beliefs. Data will be pseudonymised using the customer ID to enable data linkage. Customer email addresses are not required to enter the VRE and may be deleted from the Gateway and

	secure N-drive folder once all participant contact has been undertaken.
What is the legal basis for processing personal information? (See note e.)	Data will be processed in line with the terms in the data agreement signed by both LIDA and Sainsbury's. All data relating to individual study participants will be gained on the basis of informed consent. Sainsbury's will undertake initial contact. The consent form is split into several statements (see appendix 1 in study protocol) to ensure that the participant understands and agrees to all elements of the research. Participants must agree to all elements in the consent form in order to progress to the baseline survey and sign up for the study. Email addresses will only be shared with the research team by explicit permission of the individual participant, without this they will not be contacted further in relation to the study. Participants may withdraw at any time during data collection by emailing the lead investigator directly or by indicating their desire to withdraw in the online dietary assessment tool.
What are the plans for information sharing and access? (See note e., f., g.)	Consenting participant email addresses and loyalty card numbers will be downloaded from the online survey tool by the researcher (via the Gateway). This list of loyalty card numbers will then be shared by the IRC team with Sainsbury's via Biscom secure FTP in order for the retailer to identify their relevant customer details and transaction data. This information will then be shared with the researcher via SFTP, directly into the VRE. Due to the personal sensitivity and commercial sensitivity of the project data, there is no intention to make the data accessible to the public. However, results will be published in academic journal articles in aggregate form.
What are your plans for archiving the information at the end of the research? (See note g.)	The sensitivity of the data prevents it from being stored in an open access data archive once the project has finished. The data should be retained for a minimum of 5 years after the completion of the PhD research (final deadline September 2022). The project team are happy to take guidance from the IRC team regarding the best solution for secure data archiving needs; this may include archiving in an encrypted vault which can only be accessed by approved LIDA personnel.

What are your main data challenges? Who can help? (See note h.)	The main challenge for this project is finding suitable online platforms to carry out primary data collection. As the data partner, Sainsbury's has stipulated the requirement that any online third party survey tool must provide evidence of penetration testing in order to be signed off for use with this project. At present, the research team does not have a suitable solution in place but are seeking options for online dietary recall and survey providers which meet this requirement. Any support which the University can provide to carry out penetration testing would be very helpful.
Who is responsible for managing the information? What resources will you need? (See note i.)	The project requires coordination between the lead researcher, staff in the data team at Sainsbury's and members of the IRC team in LIDA to ensure that data is collected, transferred and made available within the VRE as required. In order to ensure that the project stays on track with its timeline, a Gantt chart will be shared with all parties, outlining the responsibilities for collecting and transferring data.
If a data breach occurs, who does the University need to inform? (See note e.)	If a data breach should occur, the following parties should be informed; Victoria Jenneson, (lead investigator) Michelle Morris, (PhD supervisor) Becky Shute, (lead project contact at Sainsbury's)
When will you notify the Data Services team that the project is drawing to a conclusion and the research environment can be closed down? (See note j.)	The data collection phase of the project is due to end in November 2020, after which point it is unlikely that any more data will need to enter the VRE. The deadline for hand in of this PhD research is September 2022, therefore it is important that the VRE remains live until this point in order to carry out analysis and aid writing up. The majority of analysis should therefore have been completed by September 2022. After completion of the PhD, the data should be securely archived for 5 years to ensure there is sufficient time to publish any additional papers and respond to any challenges which may result from PhD publications. The research team request that the IRC team liaise with them at the end of the project (September 2022) to discuss the needs for data archiving. Email addresses may be deleted much earlier in the project, once data linkage has been carried out and no further participant contact is required. The lead researcher will work with the IRC team and advise when this is suitable, at which point, a certificate of destruction is requested to be shared with Sainsbury's.

Please confirm you have provided to the Data Services Team your Project Proposal, Data Sharing Agreements and other approval letters. **Also confirm any security standards that have been requested**	**Project Proposal Yes** / ~~No~~ Filename : [STRIDE_protocol_VJ_v6.docx] **Data Sharing Agreement Yes** ~~/ No~~ Filename : [2018.01.16 LIDA Data Licence Agreement- signed by Sainsbury's LIDA/pdf; STRIDE_Data Licence Agreement APPENDIX 2.docx] Appendix 2 pending sign off from retail data partner **Ethics Approval Yes** / ~~No / Not Needed~~ Filename:[STRIDE_Ethics_V2+MM.doc] Pending, submitted 17.06.2019 **HRES approval** ~~Yes / No /~~ **Not Needed** Filename : [ ] **HRA CAG Approval** ~~Yes / No~~ / **Not Needed** Filename : [ ] **Other Approvals** ~~Yes / No~~ / **Not Needed** Filename : [ ] **NHS DPST** Not Needed **Cyber Essentials (+)** Yes / No / Not Needed **ISO27001** Yes / No / Not Needed

B.2 STRIDE ethical review form



UNIVERSITY OF LEEDS

University Research Ethics Committee - application for ethical review

Please email your completed application form along with any relevant supporting documents to ResearchEthics@leeds.ac.uk (or to FMHUniEthics@leeds.ac.uk if you are based in the Faculty of Medicine and Health) at least 6 weeks before the research/ fieldwork is due to start. Dentistry and Psychology applicants should follow their School's procedures for submitting an application.

Ethics reference (leave blank if unknown)	Student number (if a student application)	Grant reference (if externally funded)	Module code (if applicable)
	200549063		

Faculty or School Research Ethics Committee to review the application (put a 'X' next to your choice)		Arts, Humanities and Cultures (PVAR)
		Biological Science (BIOSCI)
	X	ESSL, Environment and LUBS (AREA)
		MaPS and Engineering (MEEC)
		School of Dentistry (DREC)
		School of Healthcare (SHREC)
		School of Medicine (SoMREC)
		School of Psychology (SoPREC)

Indicate what type of ethical review you are applying for:	X	Student project (PhD, Masters or Undergraduate)
		Staff project (externally or internally funded)

Section 1: Basic project details			
1.1 Research title	STRIDE; Supermarket Transaction Records In Dietary Evaluation		
1.2 Research start date (dd/mm/yy)	Proposed fieldwork start date (dd/mm/yy)	Proposed fieldwork end date (dd/mm/yy)	Research end date (dd/mm/yy)
01.10.17	01.09.19	31.11.20	30.09.21

Yes	No	
X		1.3 I confirm that I have read and understood the current version of the University of Leeds Research Ethics Policy. *The Policy is available at http://ris.leeds.ac.uk/ResearchEthicsPolicies.*
X		1.4 I confirm that I have read and understood the current version of the University of Leeds Research Data Management Policy. *The policy is available at https://library.leeds.ac.uk/info/14062/research_data_management/68/research_data_management_policy.*
X		1.5 I confirm that I have read and understood the current version of the University of Leeds Information Protection Policy. *The policy is available at http://it.leeds.ac.uk/info/116/policies/249/information_protection_policy*
X		1.6 I confirm that NHS ethical review is not required for this project. *Refer to http://ris.leeds.ac.uk/NHSethicalreview for guidance in identifying circumstances which require NHS review*

Version 1.7

Page 1 of 13

	X	1.7 Will the research involve NHS staff recruited as potential research participants (by virtue of their professional role) or NHS premises/ facilities? *Please note: If yes, NHS R&D management permission or local management permission may also be needed. Refer to http://ris.leeds.ac.uk/NHSethicalreview.*

Section 2: Contact details	
2.1 Name of applicant	Mrs Victoria Jenneson
2.2 Position (eg PI, Co-I, RA, student)	PhD student
2.3 Department/ School	School of Geography
2 4 Faculty	Environment
2.5 Work address (usually at the University of Leeds)	Leeds Institute for Data Analytics, Level 11 Worsley Building, Clarendon Way, Leeds, LS2 9JT
2.6 Telephone number	07857935495
2.7 University of Leeds email address	fs10vl@leeds.ac.uk

Section 3: Summary of the research

3.1 In plain English provide a brief summary of the aims and objectives of the research. (max 300 words). The summary should briefly describe

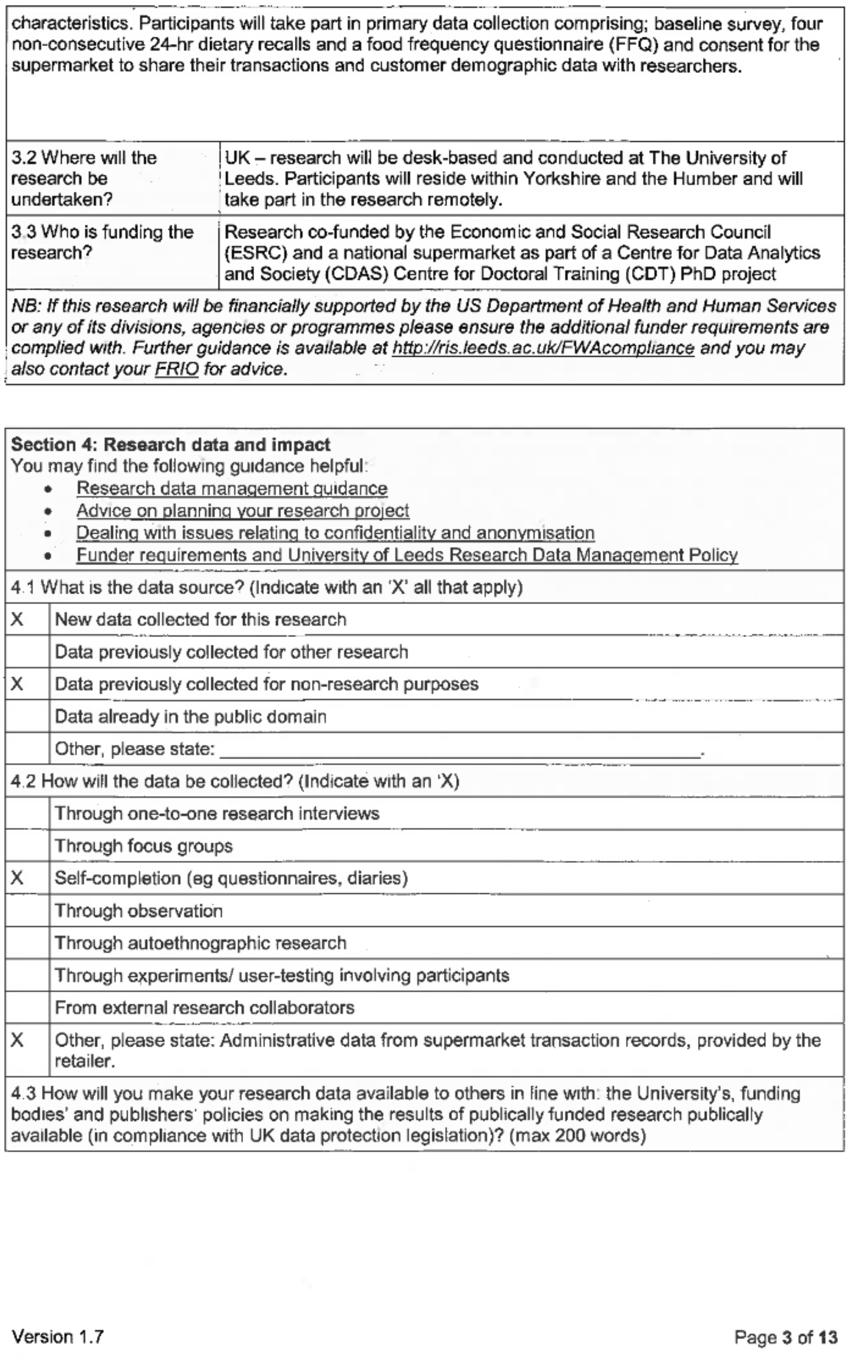
- the background to the research and why it is important,
- the questions it will answer and potential benefits,
- the study design and what is involved for participants.

Your answers should be easily understood by someone who is not experienced in the field you are researching, (eg a member of the public) - otherwise it may be returned to you. Where technical terms are used they should be explained. Any acronyms not generally known should be described in full.

Knowledge of population diets is important for designing and implementing effective population-level public health initiatives to promote adequate nutrition and prevent obesity. Currently, our understanding of UK diets comes predominantly from self-reported consumption information using traditional dietary assessment methodologies. These methods are prone to reporting biases; including systematic under-reporting of unhealthy foods and over-reporting of healthy foods, and recruitment biases; people who participate in dietary research tend to be those with an interest in the subject and with generally 'healthier' diets. Furthermore, the costs and burden for both participants and researchers restricts the sample sizes and temporality of data collection, hindering our knowledge of dietary variation by subgroup, geographic area and temporally.

Supermarket loyalty card transactions contain a vast amount of information on food and beverage purchases for millions of UK households. It is proposed that this purchase information, when linked with customer demographic data, can provide a useful proxy for dietary consumption at the population level. Furthermore, supermarket transaction data would enable detailed spatio-temporal and subgroup analyses on a scale not currently possible with traditional self-reported methods. However, in order to understand how supermarket transactions can contribute to dietary public health knowledge, we need to know more about how dietary information from transaction records compares with traditional self-reported dietary data.

This validation study will assess the statistical agreement between electronically captured supermarket loyalty card transactions and repeated self-reported 24-hour recall as a measure of population diet, among loyalty card holders from a national UK supermarket. Sensitivity analyses will be conducted to understand if the variation between methods differs according to demographic



characteristics. Participants will take part in primary data collection comprising; baseline survey, four non-consecutive 24-hr dietary recalls and a food frequency questionnaire (FFQ) and consent for the supermarket to share their transactions and customer demographic data with researchers.

3.2 Where will the research be undertaken?	UK – research will be desk-based and conducted at The University of Leeds. Participants will reside within Yorkshire and the Humber and will take part in the research remotely.
3.3 Who is funding the research?	Research co-funded by the Economic and Social Research Council (ESRC) and a national supermarket as part of a Centre for Data Analytics and Society (CDAS) Centre for Doctoral Training (CDT) PhD project

NB: If this research will be financially supported by the US Department of Health and Human Services or any of its divisions, agencies or programmes please ensure the additional funder requirements are complied with. Further guidance is available at http://ris.leeds.ac.uk/FWAcompliance and you may also contact your FRIO for advice.

Section 4: Research data and impact

You may find the following guidance helpful:

- Research data management guidance
- Advice on planning your research project
- Dealing with issues relating to confidentiality and anonymisation
- Funder requirements and University of Leeds Research Data Management Policy

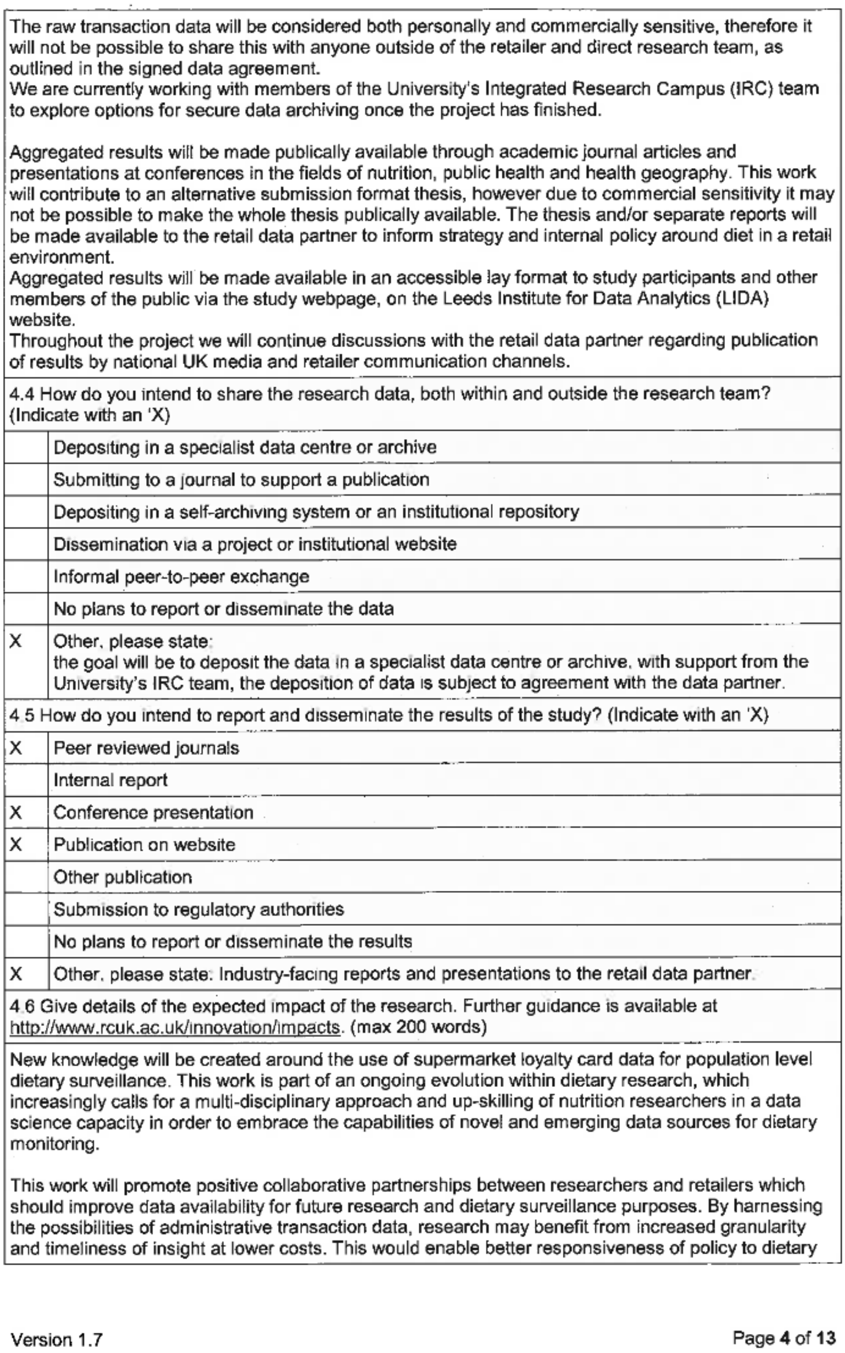
4.1 What is the data source? (Indicate with an 'X' all that apply)

X	New data collected for this research
	Data previously collected for other research
X	Data previously collected for non-research purposes
	Data already in the public domain
	Other, please state: ______________________________.

4.2 How will the data be collected? (Indicate with an 'X)

	Through one-to-one research interviews
	Through focus groups
X	Self-completion (eg questionnaires, diaries)
	Through observation
	Through autoethnographic research
	Through experiments/ user-testing involving participants
	From external research collaborators
X	Other, please state: Administrative data from supermarket transaction records, provided by the retailer.

4.3 How will you make your research data available to others in line with: the University's, funding bodies' and publishers' policies on making the results of publically funded research publically available (in compliance with UK data protection legislation)? (max 200 words)



The raw transaction data will be considered both personally and commercially sensitive, therefore it will not be possible to share this with anyone outside of the retailer and direct research team, as outlined in the signed data agreement.
We are currently working with members of the University's Integrated Research Campus (IRC) team to explore options for secure data archiving once the project has finished.

Aggregated results will be made publically available through academic journal articles and presentations at conferences in the fields of nutrition, public health and health geography. This work will contribute to an alternative submission format thesis, however due to commercial sensitivity it may not be possible to make the whole thesis publically available. The thesis and/or separate reports will be made available to the retail data partner to inform strategy and internal policy around diet in a retail environment.
Aggregated results will be made available in an accessible lay format to study participants and other members of the public via the study webpage, on the Leeds Institute for Data Analytics (LIDA) website.
Throughout the project we will continue discussions with the retail data partner regarding publication of results by national UK media and retailer communication channels.

4.4 How do you intend to share the research data, both within and outside the research team? (Indicate with an 'X)

	Depositing in a specialist data centre or archive
	Submitting to a journal to support a publication
	Depositing in a self-archiving system or an institutional repository
	Dissemination via a project or institutional website
	Informal peer-to-peer exchange
	No plans to report or disseminate the data
X	Other, please state: the goal will be to deposit the data in a specialist data centre or archive, with support from the University's IRC team, the deposition of data is subject to agreement with the data partner.

4.5 How do you intend to report and disseminate the results of the study? (Indicate with an 'X)

X	Peer reviewed journals
	Internal report
X	Conference presentation
X	Publication on website
	Other publication
	Submission to regulatory authorities
	No plans to report or disseminate the results
X	Other, please state: Industry-facing reports and presentations to the retail data partner.

4.6 Give details of the expected impact of the research. Further guidance is available at http://www.rcuk.ac.uk/innovation/impacts. (max 200 words)

New knowledge will be created around the use of supermarket loyalty card data for population level dietary surveillance. This work is part of an ongoing evolution within dietary research, which increasingly calls for a multi-disciplinary approach and up-skilling of nutrition researchers in a data science capacity in order to embrace the capabilities of novel and emerging data sources for dietary monitoring.

This work will promote positive collaborative partnerships between researchers and retailers which should improve data availability for future research and dietary surveillance purposes. By harnessing the possibilities of administrative transaction data, research may benefit from increased granularity and timeliness of insight at lower costs. This would enable better responsiveness of policy to dietary change, improving the targeting and design of strategies to improve nutrition and health of the population.

The retailer will gain insights into the dietary patters of their customer-base, enabling them to take responsible strategic action to improve the nutritive quality, availability and affordability of their offering, balancing profitability with equity and improvements in population dietary choices.

The research will highlight dietary disparities across different geographic areas and demographic groups, informing wider regional and national obesity prevention strategies.

Section 5: Protocols		
Which protocols will be complied with? (Indicate with an 'X'). There may be circumstances where it makes sense not to comply with a protocol, this is fine but should be clarified in your application.	X	Data protection, anonymisation and storage and sharing of research data
	X	Informed consent
		Verbal consent
	X	Reimbursement of research participants
		Low risk observation

Section 6: Additional ethical issues	
6.1 Indicate with an 'X' in the left-hand column whether the research involves any of the following:	
X	Discussion of sensitive topics, or topics that could be considered sensitive
X	Prolonged or frequent participant involvement
	Potential for adverse environmental impact
	The possibility of harm to participants or others (including the researcher(s))
	Participants taking part in the research without their knowledge and consent (eg covert observation of people in non-public places)
	The use of drugs, placebos or invasive, intrusive or potentially harmful procedures of any kind
	Food substances or drinks being given to participants (other than refreshments)
	Vitamins or any related substances being given to participants
	Acellular blood, urine or tissue samples obtained from participants (ie no NHS requirement)
	Members of the public in a research capacity (participant research)
	Participants who are particularly vulnerable (eg children, people with learning disabilities, offenders)
	People who are unable to give their own informed consent
	Researcher(s) in a position of authority over participants, eg as employers, lecturers, teachers or family members
	Financial inducements (other than reasonable expenses and compensation for time) being offered to participants
X	Cooperation of an intermediary to gain access to research participants or material (eg head teachers, prison governors, chief executives)
X	Potential conflicts of interest
	Internet participants or other visual/ vocal methods where participants may be identified
	Scope for incidental findings, ie unplanned additional findings or concerns for the safety or

	wellbeing of participants.
	The sharing of data or confidential information beyond the initial consent given
	Translators or interpreters
	Research conducted outside the UK
	An international collaborator
	The transfer of data outside the European Economic Area
X	Third parties collecting data
X	Other ethical clearances or permissions

6 2 For the ethical issues indicated in 6.1 provide details of any additional ethical issues the research may involve and explain how these issues will be addressed (max 200 words)

1. **Sensitive topics:** primary data will be accessible only by authorized researchers in a secure environment and not shared with the retailer for marketing purposes.
2. **Prolonged/frequent participant involvement:** primary data will be collected over 4-6-weeks; participants will complete questionnaires and 4 x 24-hr dietary recalls online to minimize participant burden.
3. **Cooperation of intermediary:** staff at the retailer will mediate contact with participants and data access. The research team will design the sampling frame and all project-related communications with participants will be agreed up-front by both parties.
4. **Conflicts of interest:** retailer corporate social responsibility (CSR) must be balanced against commitments to shareholders and profitability, which will be addressed by the retailer's own ethical review processes. The researchers will have responsibility over the design of the research project.
5. **Third parties collecting data:** baseline data will be collected through Online Surveys and dietary recalls via online dietary assessment tool (myfood24 or Oxford WebQ; dependent on the outcome of penetration testing); both platforms are widely used in research at the University of Leeds and other institutions. Once data has been downloaded and saved in a secure environment it will be deleted from third party online storage.
6. **Other ethical clearances or permissions:** the project must meet the requirements of the retailer's internal Data Clinics process which protects customer interests.

Section 7: Recruitment and consent process
For guidance refer to http.//ris.leeds.ac.uk/InvolvingResearchParticipants and the research ethics protocols.

7.1 State approximately how much data and/ or how many participants are going to be involved.

Approximately 1,800 participants are expected to complete the study and be eligible for inclusion in the final analyses. In order to achieve this sample size, it is expected that 45,000 customers will need to be contacted by the retailer.
Each participant will complete up to six online surveys as well as consent to their transaction data for 1 year during the study and 1 year prior to their participation in the study to be shared with the research team.

7.2 How was that number of participants decided upon? (max 200 words)
Please note: The number of participants should be sufficient to achieve worthwhile results but should not be so high as to involve unnecessary recruitment and burdens for participants. This is especially pertinent in research which involves an element of risk. Describe here how many participants will be recruited, and whether this will be enough to answer the research question. If you have received formal statistical advice then please indicate so here, and describe that advice.

This sample size provides adequate power for assessing statistical agreement between methods, across different subgroups under the Bland-Altman method; a minimum of 200 participants are required in the smallest subgroup. The sample for initial contact accounts for an attrition rate of 20% (based on similar research) and an expected sign up rate of 5% (based on previous market research conducted by the retailer).

7.3 How are the participants and/ or data going to be selected? List the inclusion and exclusion criterial. (max 200 words)

Participants will be selected from a database of loyalty card holders held by the supermarket. Eligible participants must reside in Yorkshire and the Humber, have an email address on file with the supermarket and be at least 18 years old. In order to be selected for this research, customers must have indicated previously that they are happy to be contacted by the retailer in relation to participation in research.

The retailer will use transaction data from the previous year to select a sample of 'typical' customers; those customers who meet a specified degree of loyalty (based on the recency, frequency and monetary value of their purchases) and buy from a broad range of food categories, suggesting that data from the retailer may be fairly representative of their overall diet.

7.4 For each type of methodology, describe the process by which you will obtain and document freely given informed consent for the collection, use and reuse of the research data. Explain the storage arrangements for the signed consent forms.

Guidance is available at http://ris.leeds.ac.uk/InvolvingResearchParticipants. The relevant documents (information sheet and consent form) need to be attached to the end of this application. If you are not using an information sheet and/ or seeking written consent, please provide an explanation.

Informed consent will be gained electronically via Online Surveys at the start of the study. Participants will receive an initial contact email from the retailer, containing a link which will take them to the participant information on the study website. Once they have read this, they may complete the consent form and baseline survey, where they will provide their email address and loyalty card number for the researchers to use for contact and data linkage respectively. Tick boxes provide an online substitute for participant signature. This one-off consent will cover all aspects of the research project and will be broken down into sections to ease participant understanding. The participant may only progress to the baseline questionnaire section of the survey if they agree with all of the statements in the consent form.

The researcher will download the completed consent forms from Online Surveys as a csv file which will be stored in a password protected project folder on the University's N-drive. This file will contain a list of consented participants eligible for further contact during the study. This list will be kept up to date in subsequent tabs to indicate if participants withdraw later in the study so that the researcher knows not to contact them further.

7.5 Describe the arrangements for withdrawal from participation and withdrawal of data/ tissue.

Please note: It should be made clear to participants in advance if there is a point after which they will not be able to withdraw their data. See also http://ris.leeds.ac.uk/ResearchDataManagement. (max 200 words)

A participant may withdraw at any time during the study period by contacting the lead researcher via email or by indicating within the myfood24/Oxford WebQ platform that they wish to withdraw. They will then be removed from all further correspondence lists and not contacted again with regards to the study. Any data already held for them will be identified and removed from the research. However, this will not be possible once data has been linked and anonymized. Participants will be informed of the cut off point for withdrawing their data from the study and reminded of the withdrawal process at each data collection opportunity.

If a participant does not complete their 24-hr dietary recall within a given time, they will receive an automatic reminder. Incomplete dietary recalls will be classed as missing data and participants will be contacted again for further participation in the study unless they actively withdraw via email or within the myfood24/Oxford WebQ platform.

The retailer's customer information team will be provided with a brief explaining how to refer participants to the lead researcher for withdrawal, should participants use this channel instead.

7.6 Provide details of any incentives you are going to use and explain their purpose. (max 200 words)

Please note: Payment of participants should be ethically justified. The FREC will wish to be reassured that research participants are not being paid for taking risks or that payments are set at a level which would unduly influence participants. A clear statement should be included in the participant information sheet setting out the position on reimbursement of any expense incurred.

In order to thank participants for their time, they will receive a series of small incentives in the form of loyalty card points from the retailer. These points hold a small monetary value and may be exchanged for goods within retailer stores. Incentives will be staged throughout the study in order to maintain participant engagement and reward appropriately those who complete more aspects of the study. Participants will be made aware of the incentives upfront in the participant information. They will be aware that they will forfeit any incentives from later study stages if they choose to withdraw.

The exact amount of incentives to be rewarded is yet to be confirmed and requires additional liaison with the retailer. However, the amount will be modest yet sufficient to encourage participation and reimburse people for their time and commitment to the research.

Section 8: Data protection, confidentiality and anonymisation
Guidance is available at http://ris.leeds.ac.uk/ConfidentialityAnonymisation

8.1 How identifiable will the participants be? (Indicate with an 'X')

	Fully identifiable
X	Identity of subject protected by code numbers/ pseudonyms
	Fully anonymised
	Anonymised but potentially identifiable
	Data only in aggregated form
	Other

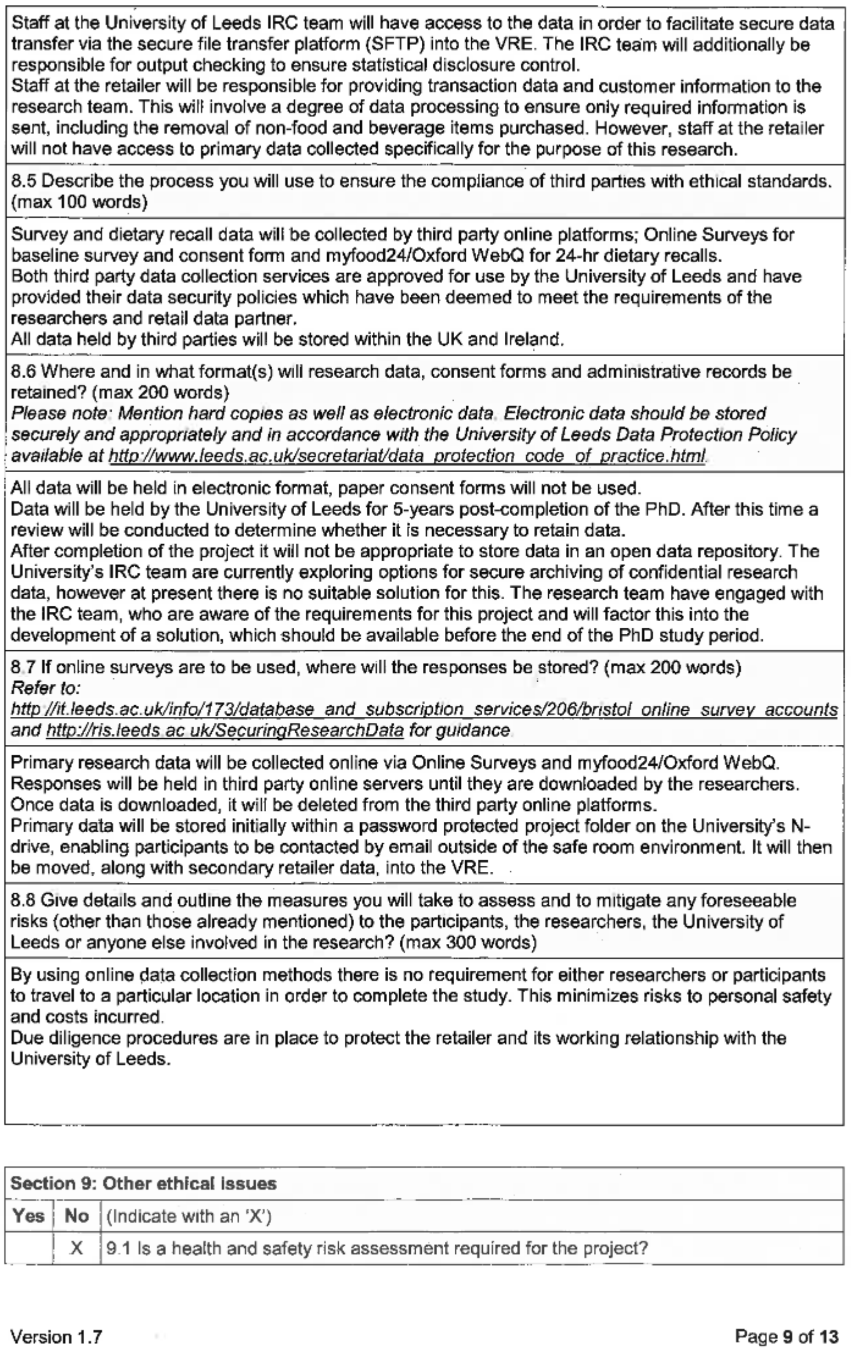
8.2 Describe the measures you will take to deal with issues of anonymity. (max 200 words)

Data collected by the research team will be linked to transaction data by the loyalty card ID number as a unique identifier.
Once data linkage has occurred, this identifier will be replaced with a pseudonymised hashed ID, after which point, the original identifier will be deleted meaning it will no longer be possible to identify participants for withdrawal.
To reduce the chance of identification, only necessary information will be collected; name, address and telephone number will not be collected.
All data collected will be stored in a secure project folder or the University's N drive, or in the IRC's Virtual Research Campus (VRE) safe room environment. Both of these environments will be password protected and only accessible by authorized project researchers.

8.3 Describe the measures you will take to deal with issues of confidentiality, including any limits to confidentiality. (Please note that research data which appears in reports or other publications is not confidential, even if it is fully anonymised. For a fuller explanation see http://ris.leeds.ac.uk/ConfidentialityAnonymisation). (max 300 words)

Data from the retailer will be transferred directly to the VRE, by the IRC data services team, and accessible only within a safe room environment. This controls all data entering and leaving the environment as there is no internet access and researchers are not permitted to take pens, paper or electronic devices into the safe room, preventing information leakage. Indirect identifiers e.g. age will also be banded.
Primary data will initially be stored in a password protected spreadsheet in a project folder on the lead researcher's M drive, before being transferred to the VRE. It will not be necessary for email addresses to enter the VRE as participants cannot be contacted from the secure environment (due to no internet connection), therefore this information will be removed before the data is transferred, reducing the identifiability of data used for linkage.
All data entering the VRE will be assessed for its suitability under the data agreement by an independent team to reduce the potential for statistical attacks (discovery of secret information through calculations by incorporating other data sources). All data leaving the VRE will also be subjected to statistical data control which, among other things, will ensure a minimum group size (n>10) for the reporting of aggregated data to reduce the chance of identifying individuals.

8.4 Who will have access to the research data apart from the research team (eg translators, authorities)? (max 100 words)



Staff at the University of Leeds IRC team will have access to the data in order to facilitate secure data transfer via the secure file transfer platform (SFTP) into the VRE. The IRC team will additionally be responsible for output checking to ensure statistical disclosure control.
Staff at the retailer will be responsible for providing transaction data and customer information to the research team. This will involve a degree of data processing to ensure only required information is sent, including the removal of non-food and beverage items purchased. However, staff at the retailer will not have access to primary data collected specifically for the purpose of this research.

8.5 Describe the process you will use to ensure the compliance of third parties with ethical standards. (max 100 words)

Survey and dietary recall data will be collected by third party online platforms; Online Surveys for baseline survey and consent form and myfood24/Oxford WebQ for 24-hr dietary recalls.
Both third party data collection services are approved for use by the University of Leeds and have provided their data security policies which have been deemed to meet the requirements of the researchers and retail data partner.
All data held by third parties will be stored within the UK and Ireland.

8.6 Where and in what format(s) will research data, consent forms and administrative records be retained? (max 200 words)
Please note: Mention hard copies as well as electronic data. Electronic data should be stored securely and appropriately and in accordance with the University of Leeds Data Protection Policy available at http://www.leeds.ac.uk/secretariat/data_protection_code_of_practice.html.

All data will be held in electronic format, paper consent forms will not be used.
Data will be held by the University of Leeds for 5-years post-completion of the PhD. After this time a review will be conducted to determine whether it is necessary to retain data.
After completion of the project it will not be appropriate to store data in an open data repository. The University's IRC team are currently exploring options for secure archiving of confidential research data, however at present there is no suitable solution for this. The research team have engaged with the IRC team, who are aware of the requirements for this project and will factor this into the development of a solution, which should be available before the end of the PhD study period.

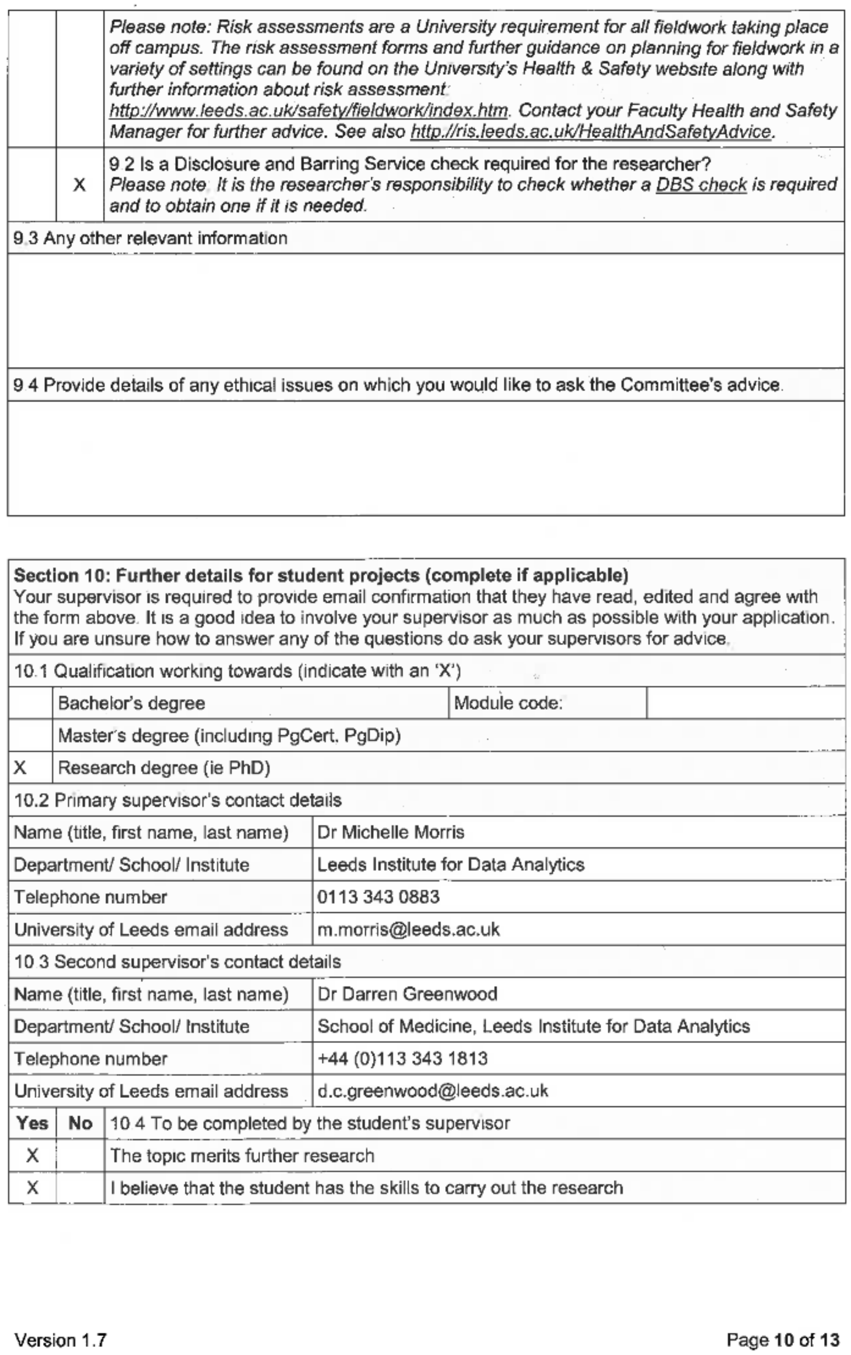
8.7 If online surveys are to be used, where will the responses be stored? (max 200 words)
Refer to:
http://it.leeds.ac.uk/info/173/database_and_subscription_services/206/bristol_online_survey_accounts and http://ris.leeds.ac.uk/SecuringResearchData for guidance.

Primary research data will be collected online via Online Surveys and myfood24/Oxford WebQ. Responses will be held in third party online servers until they are downloaded by the researchers. Once data is downloaded, it will be deleted from the third party online platforms.
Primary data will be stored initially within a password protected project folder on the University's N-drive, enabling participants to be contacted by email outside of the safe room environment. It will then be moved, along with secondary retailer data, into the VRE.

8.8 Give details and outline the measures you will take to assess and to mitigate any foreseeable risks (other than those already mentioned) to the participants, the researchers, the University of Leeds or anyone else involved in the research? (max 300 words)

By using online data collection methods there is no requirement for either researchers or participants to travel to a particular location in order to complete the study. This minimizes risks to personal safety and costs incurred.
Due diligence procedures are in place to protect the retailer and its working relationship with the University of Leeds.

Section 9: Other ethical issues		
Yes	No	(Indicate with an 'X')
	X	9.1 Is a health and safety risk assessment required for the project?



		Please note: Risk assessments are a University requirement for all fieldwork taking place off campus. The risk assessment forms and further guidance on planning for fieldwork in a variety of settings can be found on the University's Health & Safety website along with further information about risk assessment: *http://www.leeds.ac.uk/safety/fieldwork/index.htm. Contact your Faculty Health and Safety Manager for further advice. See also http://ris.leeds.ac.uk/HealthAndSafetyAdvice.*
	X	9 2 Is a Disclosure and Barring Service check required for the researcher? *Please note: It is the researcher's responsibility to check whether a DBS check is required and to obtain one if it is needed.*

9.3 Any other relevant information

9 4 Provide details of any ethical issues on which you would like to ask the Committee's advice.

Section 10: Further details for student projects (complete if applicable)
Your supervisor is required to provide email confirmation that they have read, edited and agree with the form above. It is a good idea to involve your supervisor as much as possible with your application. If you are unsure how to answer any of the questions do ask your supervisors for advice.

10.1 Qualification working towards (indicate with an 'X')

	Bachelor's degree	Module code:	
	Master's degree (including PgCert, PgDip)		
X	Research degree (ie PhD)		

10.2 Primary supervisor's contact details

Name (title, first name, last name)	Dr Michelle Morris
Department/ School/ Institute	Leeds Institute for Data Analytics
Telephone number	0113 343 0883
University of Leeds email address	m.morris@leeds.ac.uk

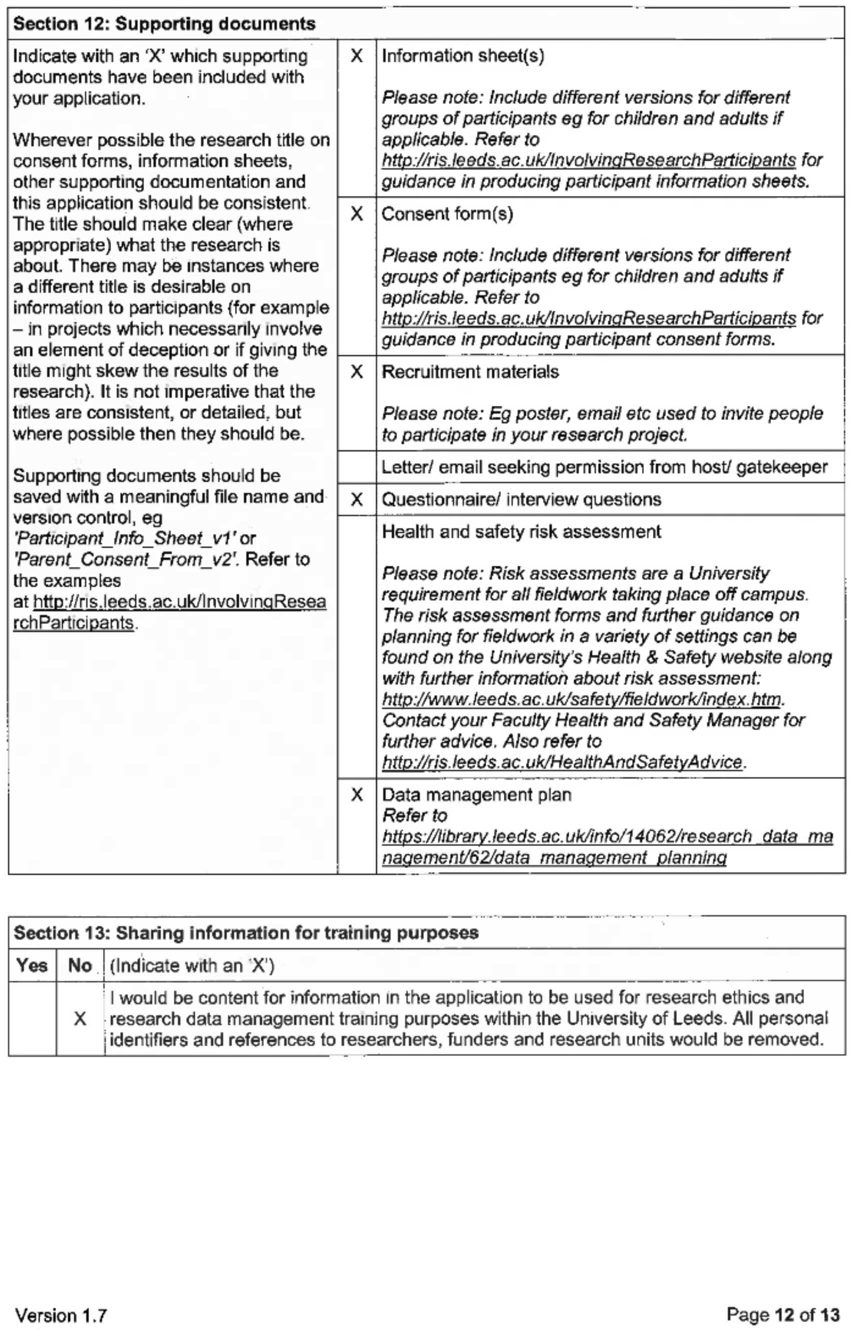
10 3 Second supervisor's contact details

Name (title, first name, last name)	Dr Darren Greenwood
Department/ School/ Institute	School of Medicine, Leeds Institute for Data Analytics
Telephone number	+44 (0)113 343 1813
University of Leeds email address	d.c.greenwood@leeds.ac.uk

Yes	No	10 4 To be completed by the student's supervisor
X		The topic merits further research
X		I believe that the student has the skills to carry out the research

Version 1.7 Page **10 of 13**

Section 11: Other members of the research team (complete if applicable)	
Name (title, first name, last name)	Professor Graham Clarke
Role (eg PI, Co-I)	Supervisor
Department/ School/ Institute	School of Geography
Telephone number	
University of Leeds email address	g.p.clarke@leeds.ac.uk
Name (title, first name, last name)	
Role (eg PI, Co-I)	
Department/ School/ Institute	
Telephone number	
University of Leeds email address	
Name (title, first name, last name)	
Role (eg PI, Co-I)	
Department/ School/ Institute	
Telephone number	
University of Leeds email address	



Section 12: Supporting documents		
Indicate with an 'X' which supporting documents have been included with your application. Wherever possible the research title on consent forms, information sheets, other supporting documentation and this application should be consistent. The title should make clear (where appropriate) what the research is about. There may be instances where a different title is desirable on information to participants (for example – in projects which necessarily involve an element of deception or if giving the title might skew the results of the research). It is not imperative that the titles are consistent, or detailed, but where possible then they should be. Supporting documents should be saved with a meaningful file name and version control, eg *'Participant_Info_Sheet_v1'* or *'Parent_Consent_From_v2'*. Refer to the examples at http://ris.leeds.ac.uk/InvolvingResearchParticipants.	X	Information sheet(s) *Please note: Include different versions for different groups of participants eg for children and adults if applicable. Refer to http://ris.leeds.ac.uk/InvolvingResearchParticipants for guidance in producing participant information sheets.*
	X	Consent form(s) *Please note: Include different versions for different groups of participants eg for children and adults if applicable. Refer to http://ris.leeds.ac.uk/InvolvingResearchParticipants for guidance in producing participant consent forms.*
	X	Recruitment materials *Please note: Eg poster, email etc used to invite people to participate in your research project.*
		Letter/ email seeking permission from host/ gatekeeper
	X	Questionnaire/ interview questions
		Health and safety risk assessment *Please note: Risk assessments are a University requirement for all fieldwork taking place off campus. The risk assessment forms and further guidance on planning for fieldwork in a variety of settings can be found on the University's Health & Safety website along with further information about risk assessment: http://www.leeds.ac.uk/safety/fieldwork/index.htm. Contact your Faculty Health and Safety Manager for further advice. Also refer to http://ris.leeds.ac.uk/HealthAndSafetyAdvice.*
	X	Data management plan *Refer to https://library.leeds.ac.uk/info/14062/research_data_management/62/data_management_planning*

Section 13: Sharing information for training purposes		
Yes	**No**	(Indicate with an 'X')
	X	I would be content for information in the application to be used for research ethics and research data management training purposes within the University of Leeds. All personal identifiers and references to researchers, funders and research units would be removed.

Section 14: Declaration

1. The information in this form is accurate to the best of my knowledge and belief and I take full responsibility for it.
2. I undertake to abide by the University's ethical and health & safety policies and guidelines, and the ethical principles underlying good practice guidelines appropriate to my discipline.
3. If the research is approved I undertake to adhere to the study protocol, the terms of this application and any conditions set out by the Research Ethics Committee.
4. I undertake to ensure that all members of the research team are aware of the ethical issues and the contents of this application form.
5. I undertake to seek an ethical opinion from the REC before implementing any amendments to the protocol.
6. I undertake to submit progress/ end of project reports if required.
7. I am aware of my responsibility to be up to date and comply with the requirements of the law and relevant guidelines relating to security and confidentiality of personal data.
8. I understand that research records/ data may be subject to inspection for audit purposes if required in future.
9. I understand that personal data about me as a researcher in this application will be held by the relevant FRECs and that this will be managed according to the principles established in the Data Protection Act.

	Applicant	Student's supervisor (if applicable)
Signature		
Name	Victoria Jenneson	Michelle Morris
Date	17.06.19	14/06/2019

B.3 STRIDE participant information

As appears on the STRIDE study page on the LIDA website https://lida.leeds.ac.uk/stride-study/

Supermarket Transaction Records In Dietary Evaluation study

What is the purpose of the research?

In most dietary research, researchers ask people to tell us what they ate, but it can be hard to remember and record everything accurately. That's why we're always looking for more effective ways to measure the diet of the UK population.

Sainsbury's, like all supermarkets, collect data about the products you buy in order to bring you the best offers tailored to you. Researchers could use this food and drink purchase data to understand how dietary habits vary across different parts of the UK and at different times of the year. This knowledge could benefit society by improving approaches to obesity prevention, which in could turn lead to cost savings for the National Health Service.

We would like to evaluate how well food and drink purchases represent what people in the UK actually eat, by comparing purchase information with more traditional self-reported dietary assessment methods.

During this study, we will link your dietary consumption data from an online Food Frequency Questionnaire, with your Sainsbury's purchases, recorded by your Nectar card. This will allow us to compare how these two pieces of information describe your diet. By doing this for lots of different people, we can understand how well purchase data might reflect the diets of the UK population. This information will help us to assess how purchase data might be used by researchers, retailers and public health practitioners in the future to help us make better food choices and live healthier lives.

Why have I been chosen?

You have been invited to take part in this research as a Nectar Card holder and based on your 2019 purchase records. Your purchase records indicate that you are a loyal Sainsbury's customer and typically buy from the majority of food categories.

What do I have to do?

1. Complete an online demographic survey

After reading all the study information and you choose to take part, please complete the online consent form and short survey. By completing the form you agree to receive study materials via email from researchers at the University of Leeds.

– When you consent to the study, you will be asked to provide your Nectar card number. This is how we will identify which purchase data belongs to you so it can be linked anonymously with your customer information and dietary recall data.

– Sainsbury's will be aware which Nectar card holders are taking part, but your participation in the study will not affect any of the usual benefits you would normally receive from Sainsbury's or Nectar.

– To maintain confidentiality, Sainsbury's will not share any information which can be used to identify you personally (e.g. your name and home address) with the University of Leeds.

– All information that you provide for the study through online surveys or your Food Frequency Questionnaire will be held by the University of Leeds. With the exception of your Nectar card number, none of the information you provide will be shared with Sainsbury's.

– The online survey will include questions about you and other members of your household, such as age and gender, which we will use alongside the information you provided when you signed up for your Nectar card.

– We will use your demographic information to put you into larger groups with other respondents who have similar characteristics (such as your ages).

– This group data will be anonymised to prevent individuals from being identified.

– This group data will be used to investigate any differences in how well purchase records represent consumption for different groups of people; for example, whether purchases of younger people are more likely to match consumption records than those of older people.

2. Complete an online Food Frequency Questionnaire

Participants will be asked to complete an online Food Frequency Questionnaire, developed by the Scottish Collaborative Group. This will take approximately 20 minutes. You will receive a link to the online Food Frequency Questionnaire. Here you should record as accurately as possible, the frequency and quantity of foods you consumed over the previous 2-3 months. To make it easier, the Food Frequency Questionnaire includes a user guide.

3. Allow your purchase records to be shared with researchers

You will only be required to participate actively in one online survey and one online Food Frequency Questionnaire. By taking part in the study you consent to share your purchase data over a longer period, however this part of the study will not take up any of your time.

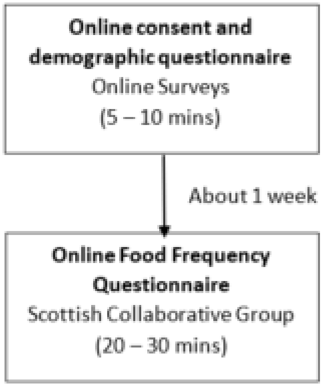
To gain a greater understanding of seasonal shopping differences and to assess whether your habits have changed as a result of taking part in the study, all food and drink purchase records linked to your loyalty card will be shared with the research team for one year during the study (2020 – 2021) and for one year prior to this (2019 – 2020).

Please remember to use your loyalty card every time you shop at Sainsbury's during the study period. We ask that you do not make any changes to your diet and food purchase habits during the study and dietary recall days, as we would like to get a sense of your usual diet.

All of the information that you provide as part of this study can be given online, at a time and place convenient for you. You will not be asked to travel to complete any aspect of this study.

Any information you provide as part of this study, the Food Frequency Questionnaires and online surveys, will be used solely for the purposes of the research described and not for marketing purposes. All items which are not food and drink will be removed from your purchase records before they are shared with the research team.

Participant Journey


Online consent and demographic questionnaire
Online Surveys
(5 – 10 mins)
About 1 week
Online Food Frequency Questionnaire
Scottish Collaborative Group
(20 – 30 mins)

What are the possible benefits of taking part?

– If you complete the study you will be entered into a prize draw for one of 5 chance to win a £75 Sainsbury's voucher*.

– You will be able to see the results of the research published on the study website, which will include links to any publications, although this may be months or even years after your participation in the research.

– Knowing that you have contributed to important research that could improve understanding of UK diets and hopefully lead to future improvements in public health.

*The prize draw will be carried out by Sainsbury's/Nectar and will take place at the end of the study in July 2021. Winners will be notified by email by Sainsbury's/Nectar. Terms and conditions for the prize draw can be found here: https://www.nectar.com/email/terms/lu_stride_ns_prize_draw?spMailingID=42552702&spUserID=ODA4MjQ3NTAyMDAyS0&spJobID=1762073048&spReportId=MTc2MjA3MzA0OAS2

What are the possible disadvantages and risks of taking part?

All the surveys and dietary recalls are designed to be as simple as possible, but participation will take up a small amount of your time.

You will also need to be able to access the internet to complete this study.

We understand that the data you provide is important to you and will be treated as confidential. There are precautionary measures in place to protect your data, including;

– only sharing data that is necessary to the research

– only allowing authorised researchers to access the data

– anonymising data prior to analysis

– using ISO accredited secure methods for transferring and storing your data, which additionally meet NHS standards.

Do I have to take part?

No, taking part is entirely voluntary and will not affect the usual communications, promotions or loyalty points you would receive from Sainsbury's or Nectar.

Can I withdraw from the study?

You may withdraw from the study at any time, until two weeks after completion of your Food Frequency Questionnaire.

You do not need to give a reason for withdrawing from the study.

If you withdraw, you will not be entered into the prize draw to win one of the 5 shopping vouchers.

If you would like to withdraw your data after having taken part, you can do so by contacting the lead investigator, Victoria Jenneson STRIDE@leeds.ac.uk.

What will happen to the results of the research project?

Results from the research will be published in academic journals, available in the public domain, presented at research conferences and will be made available to you on the study webpage. Research findings will also contribute to a PhD thesis and project report, which will be shared with Sainsbury's. Results will always be presented for groups of people to ensure individuals cannot be identified in any report or publication.

Who is organising / funding the research?

This research is being conducted by the University of Leeds in partnership with Sainsbury's Plc, as part of a co-funded Economic and Social Research Council (ESRC) PhD project.

How will my data be stored?

The anonymised research data will be stored, and the analysis conducted, within a secure research environment at the University of Leeds. This means that the data can only be accessed by authorised project researchers and the transfer of data in and out of the environment is tightly controlled. The storage and use of the research data will be conducted solely within the UK and will comply with the General Data Protection Regulation (2018), and the University's Code of Practice on Data Protection.

Data collected during this research will be stored within a restricted access University of Leeds data archive in an anonymised form for a maximum of 10 years post completion of this study. This period of data storage will enable time for all analyses to be completed and for results of the research to be published.

Your data will be used in line with the University of Leeds Research Participant Privacy Notice, available here: https://dataprotection.leeds.ac.uk/wp-content/uploads/sites/48/2019/02/Research-Privacy-Notice.pdf
The survey service provider is Online surveys, operated by Jisc. The Jisc-wide privacy notice is available here: https://www.jisc.ac.uk/website/privacy-notice.

Contact the study team

The Chief Investigators and those responsible for this study are Victoria Jenneson, PhD researcher at the Leeds Institute for Data Analytics, and Dr Michelle Morris, University Academic Fellow in Health Data Analytics, Turing Fellow, STRIDE@leeds.ac.uk, 0113-34-30883, Leeds Institute for Data Analytics, Level 11, Worsley Building, University of Leeds, Clarendon Way, Leeds, LS2 9NL.

B.4 STRIDE informed consent form

Page 2: Participant consent

Complete this page to sign up to the study.

1. I have read and understood the participant information on the STRIDE study website. *You can revisit the website any time to refresh your memory.* *Required*

○ I agree

2. I agree to participate voluntarily. I understand that I can withdraw from the study by emailing STRIDE@leeds.ac.uk *You may withdraw up to two weeks after completing your dietary questionnaire. You do not need to give a reason.* *Required*

○ I agree

3. I give permission for members of the research team to access my anonymised responses. *Your responses will be kept strictly confidential and may only be accessed by approved researchers. Your name will not be linked with the research materials, and you will not be identifiable in any reports that result from the research.* *Required*

○ I agree

4. Please provide your email address so researchers from the University of Leeds can share the dietary questionnaire with you. *Required*

Please enter a valid email address.

5. I give permission for my Nectar Card number to be used to link my survey responses with my purchase records. *Sainsbury's will be notified of your participation in the study, for your transaction data and customer information to be shared with the research team.* *Required*

○ I agree

6. Please provide your Nectar Card number (this is the last 11 digits of the long number on the front of your Nectar loyalty card).

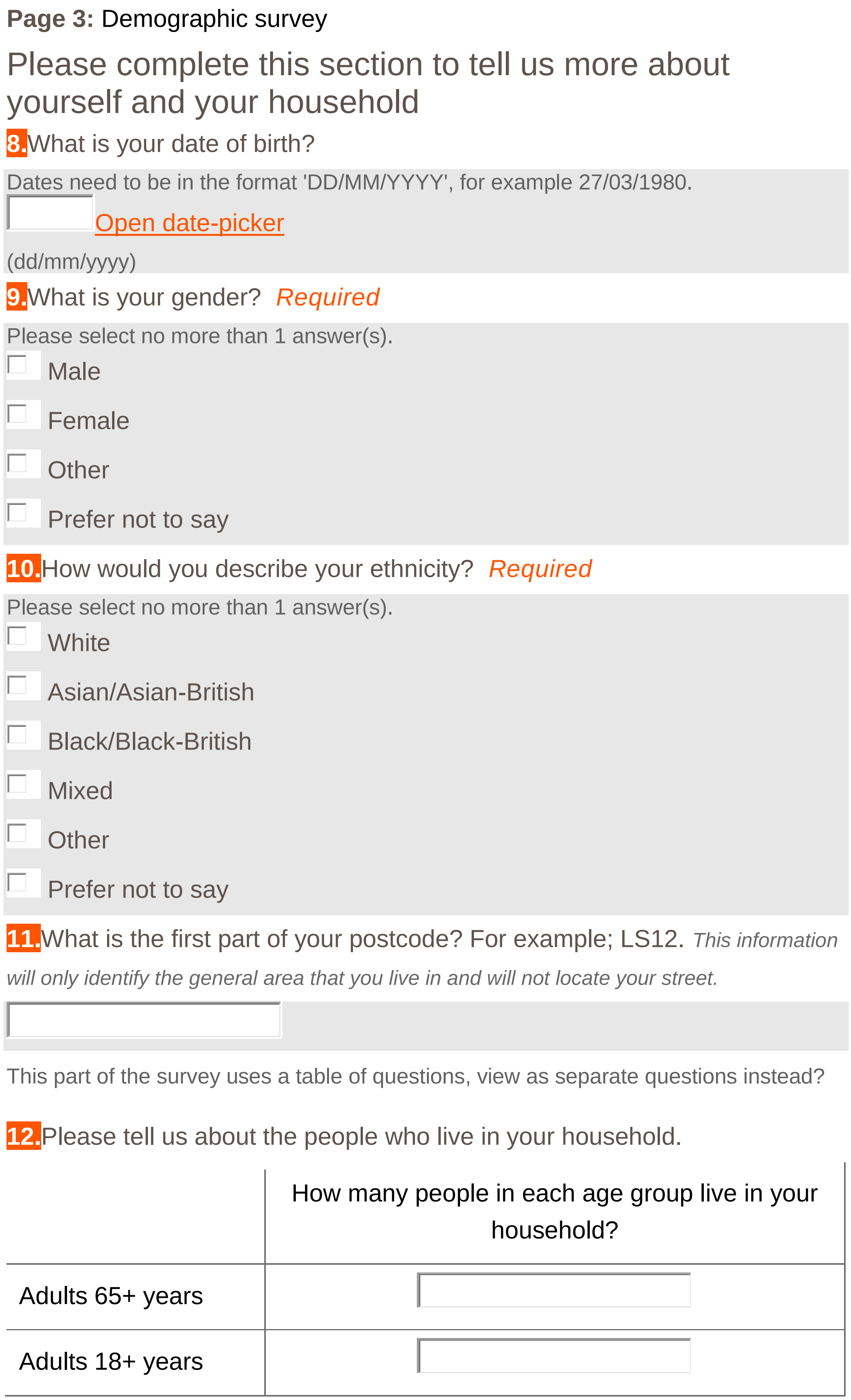
Please enter a whole number (integer).

Your answer should be no more than 11 characters long.

Thank you for giving your consent to take part in the STRIDE study.

Please continue to the next section which will contain a few questions about you and your household. It is important that you complete this section in order for you to be eligible to participate in the study.

B.5 STRIDE baseline questionnaire



Page 3: Demographic survey

Please complete this section to tell us more about yourself and your household

8.What is your date of birth?

Dates need to be in the format 'DD/MM/YYYY', for example 27/03/1980.

Open date-picker

(dd/mm/yyyy)

9.What is your gender? *Required*

Please select no more than 1 answer(s).

□ Male

□ Female

□ Other

□ Prefer not to say

10.How would you describe your ethnicity? *Required*

Please select no more than 1 answer(s).

□ White

□ Asian/Asian-British

□ Black/Black-British

□ Mixed

□ Other

□ Prefer not to say

11.What is the first part of your postcode? For example; LS12. *This information will only identify the general area that you live in and will not locate your street.*

This part of the survey uses a table of questions, view as separate questions instead?

12.Please tell us about the people who live in your household.

	How many people in each age group live in your household?
Adults 65+ years	
Adults 18+ years	

Adolescents 11-17 years	
Children 4-10 years	
Children 1-3 years	
Babies under 1 year	

13. What is your height in centimetres? *Use this handy calculator to help you convert your height in feet and inches to centimetres https://www.thecalculatorsite.com/conversions/common/height-converter.php*

Please enter a number.

14. What is your weight in kilograms? *Use this handy calculator to help you convert your weight from stones and pounds to kilograms https://www.thecalculatorsite.com/conversions/common/kg-to-stones-pounds.php*

Please enter a number.

Page 4: Your purchase habits

Please complete this section to tell us more about your current purchase habits

15. Do you consider yourself the main person responsible for grocery shopping in your household?

Please select no more than 1 answer(s).

□ Yes

□ No

□ Shared equally

This part of the survey uses a table of questions, view as separate questions instead?

16. Do you or anyone else in your household follow a special diet? Tick all that may apply

	Yourself	Someone else in your household
	Select if true	Select if true

Vegetarian	○	○
Vegan	○	○
Gluten free	○	○
Dairy free	○	○
Pescatarian (fish and vegetables)	○	○
Religious diet	○	○
Other	○	○
Not applicable	○	○

17. Excluding foods designed for pets (e.g. tinned dog food), are some of the food items you purchase intended for pets?

Please select no more than 1 answer(s).

□ Yes

□ No

18. Do you have any of the following loyalty cards? Please tick all that may apply

□ Tesco's Clubcard

□ Morrison's More

□ My Waitrose

□ Co-op Membership

□ Iceland's Bonus Card

□ Other

□ Not applicable

19. What share of all your food purchases are made in Sainsbury's stores? Please select the nearest answer.

Please select no more than 1 answer(s).

□ 0 - 20%

□ 20 - 40%

□ 40 - 60%

□ 60 - 80%

□ 80 - 100%

Page 5: Understanding dietary habits during the coronavirus (COVID-19) pandemic

Please complete this section to help us understand how your diet and purchase habits may have changed during the coronavirus (COVID-19) pandemic.

You will be asked to reflect on your habits BEFORE the COVID-19 pandemic began, and NOW.

20.BEFORE the COVID-19 pandemic began, what were the main sources of food for your household? Tick all that may apply

□ Sainsbury's (in store)

□ Sainsbury's (online)

□ Other large supermarket (in store)

□ Other large supermarket (online)

□ Convenience store (corner shop/petrol station)

□ Independent food outlet (butcher/greengrocer/bakery/fishmonger)

□ Out of home sector (restaurant/cafe/fast food/takeaway)

□ Farmers market

□ Work canteen

□ Allotment/home grown

□ Purchased by someone else

21.NOW, what are the main sources of food for your household? Tick all that may apply

□ Sainsbury's (in store)

□ Sainsbury's (online)

□ Other large supermarket (in store)

□ Other large supermarket (online)

□ Convenience store (corner shop/petrol station)

□ Independent food outlet (butcher/greengrocer/bakery/fishmonger)

□ Out of home sector (restaurant/cafe/fast food/takeaway)

□ Farmers market

□ Work canteen

□ Allotment/home grown

□ Purchased by someone else

22. BEFORE the COVID-19 pandemic began, which meals did you usually consume at home? Tick all that may apply

□ Breakfast

□ Lunch

□ Evening meal

□ Snacks

23. NOW, which meals do you usually consume at home? Tick all that may apply

□ Breakfast

□ Lunch

□ Evening meal

□ Snacks

24. BEFORE the COVID-19 pandemic began, how often did you eat meals prepared outside of the home?

Please select no more than 1 answer(s).

□ Most days

□ A few times a week

□ A few times a month

□ About monthly

□ Rarely/never

25. NOW, how often do you eat meals prepared outside of the home?

Please select no more than 1 answer(s).

□ Most days

□ A few times a week

□ A few times a month

□ About monthly

□ Rarely/never

26. Compared with BEFORE the COVID-19 pandemic began, how have your CURRENT food purchasing habits changed?

○ Not at all

○ Purchasing for more people within the household

○ Purchasing for fewer people within the household

○ Purchasing for others outside of the household

27. BEFORE the COVID-19 pandemic began, what proportion of **all food** purchased was typically wasted in your household?

Please select no more than 1 answer(s).

□ Hardly any

□ Less than 10%

□ Less than 25%

□ Less that 50%

□ More than 50%

28. NOW, what proportion of **all food** purchased is typically wasted in your household?

Please select no more than 1 answer(s).

□ Hardly any

□ Less than 10%

□ Less than 25%

□ Less that 50%

□ More than 50%

The coronavirus (COVID-19) pandemic is having an impact on all our daily lives. Many people across the UK are experiencing financial difficulties and problems accessing food as a result of the pandemic.

You can find support and advice relating to mental health and other practical issues from Mind, Gov.uk, Citizen's advice, Age UK, ACAS and more.

B.6 SCG-FFQ Fat Coding Sheet for researchers

Scottish Collaborative Group
Food Frequency Questionnaire

Coding sheet for spreads and oils

May 2014

Enter one or two codes for butter/margarine and oils/cooking fats, using the alphabetic listing attached.

If the spread does not appear on the alphabetical list but the spread or oil can be found in local shops, send information on the total, saturated, monounsaturated and polyunsaturated fat content (g/100g) to Aberdeen for coding.

If no information on fatty acid composition is obtainable, leave the coding boxes blank. In this case the main nutrient output (including total fat and fat soluble vitamins) will be calculated using code 7, but no fatty acid output will be generated for the subject.

If the subject reports not using any butter, margarine, or other spread or oil on bread, code the type of spread as 99 (no spread used)

If there is not enough information to give a code for a cooking oil, leave the coding boxes blank. The main nutrient output will be calculated using code 18 (blended vegetable oil) so that the main nutrient output is calculated, but no fatty acid data will be generated for the subject.

If the subject specifies that they do not use any cooking oil, or only use spray oil, code the type of fat or oil as 99 (no oil used).

FFQ Code	Description
1	Butter
2	Spreadable butter
3	Hard margarine (animal & vegetable fats)
4	Hard margarine (vegetable fats only)
5	Soft margarine, not polyunsaturated

6	Soft margarine, polyunsaturated
7	Blended spread, 70-80% fat
7	Fat spread, 70-80 % fat not polyunsaturated
8	Fat spread, 70% fat polyunsaturated
9	Fat spread, 60% fat, polyunsaturated
10	Fat spread, 60% fat, with olive-oil
11	Blended spread, 40% fat
12	Dairy spread, 40% fat
13	Fat spread, 40% fat, not polyunsaturated
14	Fat spread, 35-40% fat, polyunsaturated
15	Fat spread, 20-25% fat, not polyunsaturated
16	Fat spread, 20-25% fat, polyunsaturated
17	Fat spread, 5% fat
18	Blended vegetable oil
19	Corn oil
20	Olive oil
21	Peanut (groundnut) oil
22	Rapeseed oil
23	Soya oil
24	Sunflower oil
25	Compound cooking fats (solid)
26	Compound cooking fats (polyunsaturated)
27	Lard
28	Suet or beef dripping
29	Palm oil
30	Sesame oil
31	Ghee (butter-based)
32	Ghee (vegetable based)
33	Lighter/low fat spreadable butter

34	Flora Cuisine
99	No oil or spread used
99	Spray oil

Table 1.1 Butters and Margarines – Alphabetical listing by brand name

Code	Name
10	Aldi Olive spread
1	Anchor butter
7	Anchor lighter spreadable (reduced fat)
13	Anchor half fat butter
7	Anchor butter with olive oil
5	Anchor spreadable
8	Asda Best for Baking
12	Asda butter light
1	Asda English creamy butter
14	Asda light sunflower
10	Asda Olive spread
13	Asda Olive light
13	Asda Pure Gold
1	Asda smart price butter
11	Asda Smart Price reduced fat soft spread
8	Asda Soft spread
8	Asda Sunflower buttery spread
9	Asda Sunflower spread
10	Asda ‘You’d butter believe it’
	Beautifully butterly (sold in Aldi)
13	Beautifully butterly light (sold in Aldi)
	Be good to yourself (see Sainsbury’s)
10	Bertolli Olivio

13	Bertolli Olivio light
7	Bertolli with butter
10	Benecol olive spread
14	Benecol light spread
10	Benecol buttery spread
1	Butter (all kinds: Anchor/ Kerrygold/ Tesco value/ West country butter/ Somerfield English, Morrison Betta Buy, slightly salted or unsalted butter etc.)
13	Calvia (calcium enriched)
10	Carapelli
7	Clover (churned for taste/churned with less salt)
10	Casaburo Olive spread (sold in Lidl)
11	Clover light
	Clover seedburst
10	Co-op soft spread
10	Co-op buttery
1	Co-op creamery butter
10	Co-op olive
7	Co-op special blend
8	Co-op sunflower spread
1	Country life English butter/ organic butter
2	Country life organic butter
2	Country life spreadable butter
33	Country life British spreadable lighter
7	Dairygold original
3	Danpack Spreadable (sold in Lidl)
33	Danpack Spreadable lighter (sold in Lidl)
15	Delight diet
13	Delight low fat
7	Drumona spread

12	Drumona ½ fat spreadable butter
5	East End margarine
10	Easily better
3	Echo
	Filippo Berio (see Philippo Berio)
9	Flora original
16	Flora diet
9	Flora buttery
14	Flora light
16	Flora lighter than light
14	Flora pro-active low fat, lower cholesterol spread (sunflower oil)
14	Flora pro-active low cholesterol spread (olive oil)
9	Flora pro-active buttery
14	Flora omega 3
	Flora Great for baking
	Flora Gold
16	Gold (St Ivel) lowest
11	Gold (St Ivel) low fat
13	Gold (St Ivel) light
14	Gold (St Ivel) sunflower
14	Gold unsalted
7	Golden crown
10	Golden cow easispread
8	Golden churn
	Golden Sun (see Lidl)
10	Golden Sun olive gold (Lidl)
1	Graham’s churned Scottish butter
1	Graham’s organic Scottish butter
2	Graham’s spreadable butter

1	Graham’s Gold
	Healthy living (see Tesco)
10	Heavenly Buttery (sold in Lidl)
5	Herra vegetable margarine
10	I can’t believe it’s not butter
13	I can’t believe it’s not butter light
2	Kerrygold pure Irish butter spreadable
1	Kerrygold softer butter
	Kerrygold pure Irish butter
14	Kerrygold Low Low original
9	Kerrygold Low Low gold
10	Kerrygold Low Low golden cow
7	Kerrymaid
2	Kerrymaid buttery
10	Kerrymaid spread
1	Lidl Mibona butter
10	Lidl Golden sun olive gold
8	Lidl Golden sun sunflower spread
14	Lidl Golden sun sunflower 38% fat spread
2	Losely Butter
7	Lurpack Cooks - Baking
1	Lurpak salted/slightly salted/unsalted
2	Lurpak spreadable (25% vegetable oil)
7	Lurpak lighter
2	Marks and Spencers Softer butter
	Marks and Spencer lower fat spread
5	Marks and Spencer slightly salted spread
7	Marks and Spencer Touch of Butter
1	Marks and Spencer 100% entirely natural easy spreading unsalted butter

8	Marks and Spencer dairy free sunflower spread
14	Marks and Spencer lighter dairy free sunflower spread
1	Marks and Spencer freshly churned Scottish salted butter
12	Marks and Spencer half fat freshly churned butter
12	Marks and Spencer low fat butter spread
14	Marks and Spencer low fat dairy free sunflower spread
10	Marks and Spencer reduced fat olive spread
9	Marks and Spencer reduced fat spreadable (slightly salted)
2	Marks and Spencer salted butter naturally spreadable
1	Moonraker (butter)
11	Morrisons spreadable butter (low fat)
	Morrisons Olive spread
	Morrisons olive spread light
7	Morrisons soft baking spread
	Morrisons savers soft spread
	Morrisons sunflower spread
	Morrisons sunflower spread light
	Morrisons Totally Buttery
13	Morrison quality and value low fat olive
13	Morrison quality and value morning gold low fat
10	Morrison quality and value olive spread
14	Morrison quality and value reduced fat sunflower spread
1	Morrison quality and value Scottish butter
8	Morrison quality and value sunflower spread
1	Norpak
2	Norpack spreadable (sold in Aldi)
33	Norpack Spreadable lighter (sold in Aldi)
10	Olivio (see Bertolli)
10	Philippo Berio olive spread

14	Outline
1	President unsalted / salted French butter
8	Pura gold cup sunflower spread
14	Pura slimmers gold light
9	Pure dairy free soya spread
9	Pure dairy free sunflower spread
8	Pure dairy free olive spread
10	Pure organic with sunflower
	Quality and value (see Morrison)
1	Reid's dairy Scottish knight slightly salted butter
12	Rowan Glen ½ fat spreadable butter
33	Rowan Glen spreadable butter
1	Rowan Glen spreadeasy butter
14	Sainsbury's be good to yourself low fat sunflower spread
1	Sainsbury's butter (Taste the difference Normandy butter/ Taste the difference salted churned butter/ unsalted Alpine butter/ slightly salted butter)
10	Sainsbury's butterlicious
13	Sainsbury's butterlicious light
17	Sainsbury's economy reduced fat spread
10	Sainsbury's free from – dairy free
5	Sainsbury's margarine
2	Sainsbury's organic spreadable
10	Sainsbury's organic olive spread
10	Sainsbury's olive spread
13	Sainsbury's olive light
9	Sainsbury's sunflower spread
7	Sainsbury's soft spread (for baking)
	Sainsburys reduced fat soft spread
2	Sainsbury's spreadable

14	Sainsbury’s sunflower light spread
1	Scottish pride salted butter
	Solesta sunflower spread (sold in Aldi)
	Solesta sunflower spread light (sold in Aldi)
5	Silver soft
10	Somerfield butter gold
10	Somerfield olive
2	Somerfield spreadable
5	Somerfield super soft margarine
	St Ivel – see Gold/ utterly butterly
9	Stork (perfect for pastry)
10	Stork (perfect for cakes)
5	Stork (tub) – ***use as default if specific type not itemised***
5	Stork SB (special blend)
9	Stork (wrapped)
5	Tesco baking margarine
10	Tesco butter me up
13	Tesco butter me up light
13	Tesco golden light spread
13	Tesco Healthy living butter me up light
14	Tesco Healthy living enriched sunflower spread
13	Tesco Healthy living olive light spread
17	Tesco healthy living sunflower
10	Tesco olive spread
7	Tesco soft spread
2	Tesco spreadable butter
9	Tesco sunflower spread
14	Tesco sunflower light spread
17	Tesco sunflower lowest, only 5% fat

7	Utterly butterly original (St Ivel)
14	Vita D’Or (sold in Lidl)
9	Vitalite
14	Vitalite light
13	Weight Watchers Olivite
1	Yorkshire butter

You should find most codes on the above list, if not, try table 1.2 and table 1.3

Table 1.2 Butters and Margarines: Alphabetical listing by brand name – DO NOT CODE

Name	
Alsan purely vegetable margarine (palm oil base)	Meadowlea Lea Smooth
Blueband	Marks and Spencer reduced fat spread
Co-op margarine	Morrison baking margarine
Costcutters margarine	Morrison better buy far spread
Dairy crest garlic butter	Morrison better buy soft spread
Lurpark with crushed garlic	Morrison soft spread
Iceland margarine	Nuttelex
Kerry Garlic butter	Stork low fat
Kerry Low fat spread	Tesco soft spread
Meadowlea Lea Cholesterol free spread	Tesco value soft spread
Meadowlea Lea Canola spread	Tesco probiotic sunflower spread
Meadowlea Lea Milk free spread	Tomor dairy free margarine
Meadowlea Lea Hi-Omega spread	Scandinavian Style Utterly butterly (St Ivel)
Meadowlea Lea lite spread	Vegan diary free spread with soya
Meadowlea Lea Logicol spread	What, not butter?
Meadowlea Lea Logicol, Lite spread	Willow blended spread
Meadowlea Lea salt reduced spread	

Table 2.1 Oils and cooking fats: Alphabetical listing by brand name

Code	Name
22	Again and again (no cholesterol) (includes hydrogenated vegetable oils)
21	Alfa One Rice Bran Oil
24	Asda (Chinese) stir fry oil
20	Asda olive – pomace oil
24	Asda pure grapeseed oil
21	Asda groundnut oil
22	Asda Pure vegetable oil
27	Asda smart price lard
24	Asda sunflower and olive oil
24	Asda walnut oil
20	Boi Organic Costa D'Or
28	Britannia finest beef dripping
22	Canola oil
22	Carotino nature oil (vitamin rich) / mild and light cooking oil / red palm and canola oil
	Coconut oil
25	Cookeen
19	Corn oil (all kinds: Mazola/ Sainsbury's/ Tesco etc)
20	Chalice (lemon infused olive oil)
24	Chalice stir fry oil (blend of sunflower, garlic, ginger)
22	Chip Shop
27	Crisp 'n dry solid
22	Crisp n dry
24	Flax Oil (granovita organic)
34	Flora Cuisine
26	Flora white
99	Fry light

25	Frytex
31	Ghee (butter-based)
32	Ghee (vegetable-based)
21	Groundnut oil (all kinds: Chalice/ Sainsbury’s/Tesco etc)
20	KTC olive pumice oil
22	KTC mustard oil / vegetable oil
24	Lidl Golden Sun sunflower oil
19	Lidl Golden Sun frying oil
	Lurpack ‘ Cooks’ – cooking liquid
	Lurpack ‘ Cooks’ – clarified butter
99	Lurpack ‘Cooks’ – cooking mist
20	Macadamia nut oil
19	Mazola pure corn oil
24	Morrison grapeseed oil
24	Morrison stir fry oil
27	Morrison savers lard
	Morrison signature goose fat
27	Morrison finest quality lard
20	Olivado avocado oil
20	Olive oil (all kinds)
20	Olive pomace oil
18	Olivio cooking oil (vegetable oil + 15% olive oil)
18	Pura light touch
22	Pura vegetable oil
24	Morrison sunflower oil
22	Morrison vegetable oil
22	Sainsburys almond oil
	Sainsburys duck fat

	Sainsburys goose fat
27	Sainsbury’s lard
24	Sainsbury’s grapeseed
22	Sainsbury’s pure vegetable oil (rape)
24	Sainsbury’s sunolive (85% sunflower + 15% olive oil)
30	Sesame oil (toasted), (all kinds: Sainsbury’s etc)
22	Somerfield vegetable oil
23	Soyola
99	Spray oil (Tesco sunflower etc.)
25	Spry crisp and dry solid cooking oil
24	Sunflower oil (all kinds: Flora/ Sainsbury’s/ Somerfield/ Tesco/ Vita d’or)
12	Tesco half fat butter
22	Tesco pure vegetable oil
26	Trex pure vegetable fat
23	Vita d’Or vegetable oil (soya)
19	Vita d’Or corn oil
24	Walnut oil (all kinds: Chalice/ Sainsbury’s, etc)

You should find most codes on the above list, if not, try table 2.2

Table 2.2 Oils and cooking fats: Alphabetical listing by brand name – DO NOT CODE

Chalice chilli infused sunolive oil	Nisa vegetable oil
Chalice chilli infused oil made with fresh chillis	Pura vegetable lard
Chalice (garlic infused oil/basil infused oil)	Pura vegetable oil (solid)
Heart content oil	Pure additive free vegetable oil
Lidl Vita d’Or pure vegetable oil	
Loscoe chilled foods Ltd pork dripping with jelly	