Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Diet and other risk indicators associated with dental problems in Irish preschool children

A thesis submitted to the University of Dublin, Trinity College for the Degree of Doctor of Philosophy

Prepared by: Michael James Crowe B.Sc., B.Dent.Sc.

Division of Restorative Dentistry and Periodontology, Dublin Dental University Hospital, University of Dublin, Trinity College, Lincoln Place, Dublin 2

March 2018

Declaration

I declare that this thesis has not been previously submitted as an exercise for a degree at this or another university.

I declare that this report is entirely my own work unless otherwise stated, and in which case, acknowledgements and references are given to the work of others.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

_____

Michael Crowe

March 2018

Summary

Investigations into the wider bioecological understanding of dental problems in early childhood are frequently limited in national surveys. Initially, this research used Classification tree analysis (CTA) and logistic regression to explore multilevel interactions among key aspects of child and primary caregiver (PCG) psychosocial and physical health affecting dental problems in preschool children. Data were derived from the Growing Up in Ireland (GUI) study, a nationally representative sample of 9 month olds (n=11,134) in 2007/2008 followed-up at age 3 years (n=9,793) in 2010/2011. Analysis included PCG reports of childrens' dental problem visits, general health, temperament, emotional and behavioural difficulties as well as their own general health, stress and depression, relationship and sociodemographic variables. Dental problems were reported among 2.7% of infants at 9 months of age and 5.0% at 3 years. CTA identified infant temperament (ICQ unpredictable) as the primary predictor of dental problems at 9 months and child global health at 3 years of age. Key aspects of infant/child and PCG health, and psychosocial characteristics associated with reported dental problems should be considered in future multidisciplinary approaches to child health.

A data science approach was also used to investigate the dietary aspects of the GUI infant cohort at 3 years (second wave) as a poor quality diet may be a common risk factor for both obesity and dental problems such as caries. CTA was used to classify variables and describe interactions between multiple variables including socio-demographics, dietary intake, health-related behaviour, BMI and a dental problem. The CTA model showed a sensitivity of 67% and specificity of 58.5% and overall correctly classified 59% of children. Ethnicity was the most significant predictor of dental problems followed by longstanding illness or disability, mother's BMI and household income. The highest prevalence of dental problems was among children who were obese or underweight with a longstanding illness and an overweight mother. Frequency of intake of some foods showed interactions with the target variable. The prevalence of overweight or obesity in 3 year old children was approximately 23%. The common risk factor approach may be a pragmatic means of developing shared modifiable strategies for both dental and weight problems.

To further explore the pattern of cariogenic food intake data mapping protocols were developed to link dietary intake estimates for 3 year old children from the National Preschool Nutritional Survey (NPNS) and the second wave of the GUI infant cohort. The GUI survey used a short frequency questionnaire (SFQ) which provided a limited list of “healthy” and “unhealthy foods”. This limited dietary intake was augmented by unidirectional mapping from the more detailed National Preschool Nutrition Survey (NPNS) which used a detailed 4 day weighed food diary. Through mapping the food codes in this manner and estimating the degree of *non-covered* food it was possible to visualise the relative performance of the brief dietary instrument compared to the more detailed one especially in capturing specific food types, e.g., high sugar foods. The SFQ did not capture a substantial portion of habitual foods consumed by 3 year olds in Ireland. Researchers interested in focussing on specific foods, such as those high in sugar, could use this approach to easily assess the proportion of foods *covered*, *non-covered* or *partially-covered* by reference to the mapped food database.

The estimation of cariogenic food and drink (CF) intake using two different methods, in Section 6 of this analysis, highlighted how presentation and reporting can affect the interpretation of CF consumption data. Bean plots also illustrated the usefulness of visualising the overall distribution of intake when using different methods of estimation and comparing snacking and main meal consumption patterns. Key findings indicated that all children consumed CF over the NPNS 4-day period and the GUI survey covered less than half of the CF items selected in NPNS. More than one-third of all eating occasions were described as snacks which were consumed twice per day, on average. Association analysis of the meals and snacks provided an insight into the combination of meal components and how CF was consumed with other non-cariogenic foods (NCF) with biscuits, squashes, cordials, fruit juice drinks and chocolate confectionary the most commonly consumed CF snack items. Using the frequent item sets and association rules from this analysis alluvial plots visualised the CF and NCF interactions of both snacks and main meals and demonstrated the importance of understanding the pattern of CF consumption.

Finally, a free sugar algorithm was developed to determine the dietary free sugar content of foods in the NPNS database. The key sources of FS were similar to the items that do not contribute significantly to nutrient intake but were primary contributors to CF intake. Almost three-quarters of 3 year olds had FS intake greater than the WHO recommendation that FS intake is a maximum

10% of total energy intake. A crude simulation which “excluded” all FS, from snacks only, of soft drinks, confectionery, cakes and biscuits and sugar demonstrated that mean daily intake of FS would reduce from 14.1% to 11% of total energy intake. As a population level estimate, this would double the proportion of three year olds, from one-quarter to one-half, meeting the maximum WHO recommendation of 10% total energy intake from FS.

Acknowledgements

My sincere gratitude and thanks to all the wonderful people, who have made my PhD journey an interesting, enjoyable experience. In particular, I would like to thank my supervisor Michael O' Sullivan, without whose constant motivation and guidance, I would neither have started nor finished this work. Many thanks Mike, for providing me firstly, with the opportunity, and secondly with the unwavering support to complete this research, it is greatly appreciated. Aifric O' Sullivan was a wonderful co-supervisor and educator in understanding diet and nutrition. Thank you Aifric, for your clear, insightful supervision and constant willingness to go the extra mile. Oscar Cassetti, opened the world of data science to me. Anyone who makes statistics exciting has a special gift and I was fortunate to encounter a mathematician with a fascination for health and nutrition. Thank you, Oscar for helping me to learn so much from all those meetings and Skype calls across the world.

I am extremely grateful to Professor June Nunn for her encouragement in the early stages and to Professor Brian O' Connell who provided continued inspiration and support. Many thanks to the Executive Team of the Dublin Dental University Hospital for funding this research. My thanks also to Professor Colman McGrath who helped originate the research topics and assisted greatly in the early stages of the work. Lorraine Swords provided insight into the fascinating world of psychology and I am very grateful for her guidance.

Thanks to all the staff in Department 2 and especially to Jeff for insights into his golf swing on late Friday evenings. To all the surgery staff for putting up with me (even more than usual) for the last 4 years and for pretending there has been no impact at work. My thanks to Catherine Lovett for her very helpful advice on formatting.

My family also played a significant role on my PhD journey. I am indebted to:

My recently departed, wonderful and vibrant, young Aunt, Gerry. Your sudden loss has shaken our core, "death lies on her like an untimely frost". My wonderful brothers, Richie and Darragh, for constantly ensuring that "the very substance of the ambitious is merely the shadow of a dream". My Mum who was always proud of her sons' achievements and encouraged us from an early age. My recently departed Dad, who inspired a life-long love of reading and

learning, your guiding hand will remain with me always and I will cherish our memories forever.

My joyous and extraordinary children, Séan, Aisling and David, without whose never-failing advice and help about statistics and life this thesis would have been finished in half the time.

Finally, I dedicate this thesis to my loving (and patient) wife, Dolores. Bloomsday will always hold a special place in my heart.

Table of Contents

List of Figures

(Figure 5.2a) or weight (Figure 5.2b) of a food item consumed in NPNS that were *non-covered* by the mapped GUI data model.

Figures 5.2a and 5.2b Food frequency and consumption weight non-covered by GUI survey by the day of the week representing the distribution of the ratio of consumption counts* (Figure 5.3a) or weight (Figure 5.3b) of a food item consumed in NPNS that were *non-covered* by the mapped GUI data model over the total food covered.

Figures 6.1a and 6.1b Bean plots illustrating the distribution patterns (kernel density estimates) of weight of food consumed (Figure 6.1a) and frequency of consumption (Figure 6.1b) of snacks and main meals estimated using the mean daily intake method.

Figures 6.2a and 6.2b Bean plots illustrating the distribution patterns (kernel density estimates) of weight of food consumed (Figure 6.2.a) and frequency of consumption (Figure 6.2.b) of snacks and main meals estimated using the average consumption method.

Figure 6.3 Alluvial plot of association analysis of the 10 most frequent eating occasions of cariogenic food using keys* for NPNS†-derived food codes for cariogenic and non-cariogenic descriptors.

Figure 6.4 Alluvial plot illustrating association analysis of 2 eating occasions with cariogenic food using keys* for NPNS†-derived food codes for cariogenic and non-cariogenic descriptors.

Figure 7.1 Flow chart depicting the processes in the mapping and analysis of free sugar intake using the Growing Up in Ireland (GUI) database augmented with the national preschool nutrition survey (NPNS).

Figure 7.2 Snippet illustrating the free sugar mapping procedure of the NPNS food data.

Figure 7.3 Sample of dataframe in RStudio, illustrating the mapped GUI database (from NPNS) with mean total sugar content of food items (g/100g) and amount of free sugar in the meal (g).

Figure 7.4a and 7.4b comparison of distribution of free sugar estimations (g) carried out in this analysis (a) and that carried out by a previously reported mapping (b).

Figure 7.5 (a-d) Daily intake of total (a, c) and free sugar (b, d) *covered* and *non-covered* by GUI food groups for 3 year old children by amount (g/day) and as a percentage of Total Energy Intake (%TEI).

List of Tables

List of Abbreviations

BMI	Body mass index
CART	Classification and regression tree analysis
CF	Cariogenic Food and Drink
CHAID	Chi-square Automatic Interaction Detection
CTA	Classification Tree Analysis
ECC	Early Childhood Caries
EO	Eating Occasions
FD	Food Diary
FS	Free Sugars
GUI	Growing Up in Ireland
IOTF	International Obesity Task Force
KDD	Knowledge Discovery in Databases
NPNS	National Preschool Nutrition Survey
NCF	Non-Cariogenic Food and Drink
NLP	Natural Language Processing
RTEBC	Ready To Eat Breakfast Cereals
TEI	Total Energy Intake
TS	Total Sugars
WHO	World Health Organisation

Glossary

Accuracy; this is by far the most widely known measure of classifier performance. For a classifier, accuracy is defined as the number of items categorised correctly divided by the total number of items. It is simply what fraction of the time the classifier is correct. At the very least, a classifier must be accurate. In terms of a confusion matrix, accuracy is (TP+TN)/ (TP+FP+TN+FN). *Accuracy* used in a classification sense is not to be confused with *accuracy* used in a numeric sense which is defined as score-based accuracy as a numeric quantity that can be decomposed into numeric versions of trueness and precision.

Accuracy paradox; states that predictive models with a given level of accuracy may have greater predictive power than models with higher accuracy. Despite optimizing classification error rate, high accuracy models may fail to capture crucial information transfer in the classification task. This is why it is important to use parameters other than accuracy to evaluate model performance.

Algorithm; a series of repeatable steps for carrying out a certain type of task with data.

Association analysis; used to find objects or attributes that frequently occur together. The algorithms look for all the itemsets (subsets of transactions) that occur more often than in a minimum fraction of the transactions and then turn those itemsets into rules.

Bagging; stands for “Bootstrap Aggregating” whereby bootstrap samples are drawn randomly with replacement.

Bar chart; this is a histogram for discrete data: it records the frequency of every value of a categorical variable.

Big data; not well defined and generally refers to the three V’s- volume, variety and velocity. This has now broadened to include technologies and decisions/solutions for problems. In clinical research ‘Big Data’ generally refers to data on a large number of variables per person or data in a large number of persons.

Binning; data binning is a data pre-processing technique used to reduce the effects of minor observation errors. Statistical data binning is a way to group a number of more or less continuous values into a smaller number of "bins". Histograms are an example of data binning.

Boosting; machine learning algorithms that try to improve the accuracy of a classifier by a reweighting of misclassified samples.

Black-box; many machine learning applications involve predictive analytic modelling using black-box techniques. Breiman (2001) stated- “the analysis in this culture considers the inside of the (black) box complex and unknown”. The approach is to find a function or algorithm that operate on independent variables to predict the response variables.

Bootstrap; a popular method for variance estimation in surveys. Data are resampled repeatedly, and a statistic is calculated for each resampling to form an empirical distribution for that statistic.

Bonferroni adjustment; the Bonferroni correction adjusts probability (p) values because of the increased risk of a type I error when making multiple statistical tests.

Code fragment; also described as code snippet, is a piece of executable code.

Code chunks, can be used to render R output into documents or to display code for illustration.

Cross-validation; a widely used method for estimating prediction error for a model, Ideally, if there is enough data, we would set aside a validation set would be set aside and used to assess the performance of the prediction model. Since data are often scarce, this is usually not possible. In this situation, a K-fold cross- validation can use part of the available data to fit the model, and a different part to test it.

Data.frame; the main data type in R, similar to a spreadsheet with tabular columns and rows.

Data.table; a high speed extension of data.frames.

Data mapping; in data management, data mapping is the process of creating data elemant mappings between two distinct data models. Data mapping is used as a first step for a wide variety of data integration tasks.

Data Science; the merging sciences of statistics, machine learning, computer engineering and visualisation.

Data mining; data mining (DM) is the core of the KDD process, involving the inferring of algorithms that explore the data, develop the model and discover previously unknown pattern.

Decision tree; modern technique for performing nonlinear regression or classification by iteratively splitting predictors.

Dependent variable; see output.

Determinant: an attribute or exposure that increases the probability of occurrence of a specified outcome.

Dispersion; refers to the variation within a sample or a population and standard measures include the variance and the range.

Eating occasions: all snacks and main meals were collectively described as eating occasions.

Fitted model; a fitted model is just the closest model from a family of models. That implies that one has the “best” model (according to some criteria); it doesn’t imply that model is “good” and it certainly doesn’t imply that the model is “true”.

Greedy (algorithm); a programming technique which always seems to make the locally optimal choice at each stage in the hope of finding a global optimum.

Histogram; an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable. Histograms display the distribution of a continuous variable by dividing up the range of scores into a specified number of bins on the x-axis and displaying the frequency of scores in each bin on the y-axis. A basic histogram bins a variable into fixed-width buckets and returns the number of data points that falls into each bucket.

Independent variable; see input.

Input; often called predictor or independent variable in statistical literature, also called features in the pattern recognition literature.

Kernel density plot; also known as density plot, kernel density plots can be an effective way to view the distribution of a continuous variable and is based on

a nonparametric method for estimating the probability density function of a random variable.

K-fold cross-validation; practice of splitting the derivation data into K equal parts. The model is then trained on K-1 parts and validated on the remaining part.

Knitr; modern package for interweaving R code with Markdown.

Knowledge Discovery in Databases (KDD); an automatic, exploratory analysis and modelling of large data repositories. KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets.

LMS parameters; the LMS parameters are the power in the Box-Cox transformation (L), the median (M) and the generalised coefficient of variation (S).

Markdown; simplified formatting syntax used to produce elegant HTML documents in simple fashion.

Model; a statistical model makes assumptions about the generation of sample data and embodies an expectation of the relationships between the data and various factors in the actual population data.

Machine learning; modern, computationally heavy statistics, set of tools ranging from artificial neural networks to random forests and decision trees. Used in 'data mining', knowledge discovery in databases' and pattern recognition.

Non-parametric model; a model where the response does not necessarily follow the regular GLM distributions such as Normal, Logistic or Poisson

Outcome criteria; output variables used to rank or measure the desirability or undesirability of possible model outcomes. Their values are determined by the input quantities and the models that use them.

Output variable; also called outcome, target, response or dependant variable

Overfitting; an analysis that corresponds too exactly to a particular set of data and may therefore fail to fit additional data or predict future observations reliably. Essentially, an overfit model appears promising on the training data and performs poorly on new data.

Probability distribution; the probability structure of a random variable, say *y*, is

described by its probability distribution. If *y* is discrete, often called the probability distribution of *y*, say *p*(*y*), the probability mass function of *y*. If *y* is continuous, the probability distribution of *y*, say *f*(*y*), is often called the probability density function for *y*.

Random error; also known as variability, random variation, or 'noise in the system'. The heterogeneity in the human population leads to relatively large random variation in clinical trials. Random error corresponds to imprecision.

Recursive partitioning; where the feature space is recursively split into regions containing observations with similar response values. Decision trees are an example of recursive partitioning. As each sub-population may in turn be split an indefinite number of times until the splitting terminates it is described as recursive.

Risk indicator; an exposure that is associated with an outcome only in cross-sectional data. A risk indicator may be a probable, or putative, risk factor.

Risk factor; Risk factor: an environmental, behavioural, or biologic factor confirmed by temporal sequence, usually in longitudinal studies, which if present directly increases the probability of a disease occurring, and if absent or removed reduces the probability.

Risk marker; an attribute or exposure that is associated with increased probability of disease but is not necessarily a causal factor. Sampling techniques include random oversampling and random undersampling. Oversampling randomly duplicates the minority class samples, while undersampling randomly discards the majority class samples in order to modify the class distribution.

Sensitivity; also called recall or True positive ratio (TPR) = TP/TP+FN. Calculates proportion of positives correctly identified.

Specificity; also called True Negative Ratio (TNR) =TN/TN+FP. Measures proportion of negatives correctly identified.

Stratification; grouping the study population into subgroups by their homogenous characteristics before sampling so as to improve the representativeness of a sample.

Supervised learning; a type of machine learning algorithm in which a system is taught to classify input into specific, known classes. For each observation of

the predictor measurements there is an associated response measurement.

Systematic error or bias; refers to deviations that are not due to chance alone. Bias corresponds to inaccuracy.

Total energy intake: the sum of all daily energy (kilojoules/kilocalories) from fat, protein, carbohydrates and alcohol consumption.

Unsupervised learning; refers mostly to techniques that group instances without a prespecified, dependent attribute. A class of machine learning algorithms designed to identify groupings of data without knowing in advance what the groups will be.

Publications

Presentations and proceedings

CROWE, M., O' SULLIVAN, M., NUNN, J.H., MCGRATH. Dental Problems in Early Childhood Are Associated with Health and Relationships. International Association for Dental Research/Pan European Regional Congress Dubrovnik, Croatia.

CROWE, M., O' SULLIVAN, M., NUNN, J.H., MCGRATH. 2014. Dental Problems During Early Childhood Associated with Caregiver and Child Health. International Association for Dental Research/Pan European Regional Congress Dubrovnik, Croatia.

CROWE, M., O' SULLIVAN, M., NUNN, J.H., MCGRATH. 2014. Maternal and Child Psychosocial Factors associated with Dental Problems During Early Childhood. Growing Up in Ireland Sixth Annual Research Conference.

CROWE, M., O' SULLIVAN, M., MCGRATH, C., CASSETTI, O. & O' SULLIVAN, A. 2015. Diet, Dental Problems and Obesity in 3 year old Children in Ireland: Classification and Regression Tree Analysis of selected independent variables. Growing Up in Ireland Seventh Annual Research Conference.

CROWE, M., O' SULLIVAN, M., CASSETTI, O. & O' SULLIVAN, A. 2016. Links between diet, dental and weight problems in 3-year-old children? Irish Division, Annual Scientific Meeting of the International Association for Dental Research. University College, Cork.

CROWE, M., O' SULLIVAN, M., CASSETTI, O. & O' SULLIVAN, A. 2016. Data mapping to augment dietary intake values from a nutritional database to a national cohort survey: protocols to improve quality of reported food intake (OCE3) doi: 10.1017/S0029665116002032. Nutrition Society Summer meeting- New technology in Nutrition research and practice, Joint British and Irish Conference, University College Dublin.

CROWE, M., O' SULLIVAN, M., CASSETTI, O. & O' SULLIVAN, A. 2016. Combining nutritional data from two surveys to augment dietary intake estimates. Growing Up in Ireland Eighth Annual Research Conference.

CROWE, M., O' SULLIVAN, M., CASSETTI, O. & O' SULLIVAN, A. 2016. Connecting diet, dental health and obesity in preschool children. Oral

presentation. UCD Childhood and Human Development Research Centre Launch, CHD-RC, University College Dublin.

CROWE, M., O' SULLIVAN, M., CASSETTI, O. & O' SULLIVAN, A. 2016. Combining data from a nutritional database to a national cohort survey using a data mapping protocol. Poster presentation, Faculty of Health Sciences Research Meeting, Trinity College, Dublin.

CROWE, M., O' SULLIVAN, M., CASSETTI, O. & O' SULLIVAN, A. 2017. Cariogenic food and drink consumption and dental problems in 3 and 5-year olds in the GUI Infant Cohort. Growing Up in Ireland Ninth Annual Research Conference.

Journal Publications

CROWE, M., O'SULLIVAN, A., MCGRATH, C., CASSETTI, O., SWORDS, L. & O'SULLIVAN, M. 2016. Early Childhood Dental Problems Classification Tree Analyses of 2 Waves of an Infant Cohort Study. *JDR Clinical & Translational Research,* 1**,** 275-284.

CROWE, M., O' SULLIVAN, M., CASSETTI, O. & O' SULLIVAN, A. 2017. Weight Status and Dental Problems in Early Childhood: Classification Tree Analysis of a National Cohort. *Dentistry Journal,* 5**,** 25.

CROWE, M., O' SULLIVAN, M., CASSETTI, O. & O' SULLIVAN, A. 2017.Data mapping protocols to augment reported food intake. Submission March 2018, *Frontiers in Nutrition, Nutrition Methodology*.

CROWE, M., O' SULLIVAN, M., CASSETTI, O. & O' SULLIVAN, A. Estimation and consumption pattern of free and total sugar intake in 3 year old children. Invited paper, in preparation for submission special issue in June 2018, *Dentistry Journal.*

Chapter 1. Introduction

1.1. General Introduction

Dental problems that occur in preschool children can have both immediate and life-long impacts on both oral and general health (US Department of Health and Human Services, 2000, Lee et al., 2013, Sheiham, 2005, Dye, 2017). While oral health issues that require a visit to the dentist vary from physiological (e.g., teething), to trauma, (e.g., tooth fracture), the most common reason in childhood is pain due to dental caries (Boeira et al., 2012, Daher et al., 2015). Early childhood caries (ECC) is the most prevalent dental problem in preschoolers (Public Health England, 2013b, Gussy et al., 2016, Wagner and Heinrich-Weltzien, 2017) and one of the most common reasons for young children requiring admission to hospital for extraction procedures under general anaesthetic (Public Health England, 2015a, Watt and Rouxel, 2012, Leong et al., 2013, Declerck et al., 2008). Increasing rates of ECC (Dye, 2017, Declerck et al., 2008), particularly among disadvantaged households (Baggio et al., 2015, Vargas and Ronzio, 2006, Newton and Bower, 2005), is concerning, especially when ECC is a preventable public health problem (Baggio et al., 2015, Vargas and Ronzio, 2006, Newton and Bower, 2005, Dye, 2017, Declerck et al., 2008) In the primary dentition dental caries affected more than 620 million children worldwide in 2010 (Kassebaum et al., 2015, Selwitz et al., 2007) and is also one of the best predictors of future caries in the permanent dentition (Tinanoff and Reisine, 2009, Gussy et al., 2006, Selwitz et al., 2007).

Infants and young children depend on the primary caregiver (PCG), usually their mother, for decisions relating to their healthcare (Hooley et al., 2012b) and the lifestyle and oral health behaviour of the PCG is strongly related to the oral health of the child during the first years (Meurman and Pienihäkkinen, 2011). However, while studies exploring risk factors for ECC in preschoolers have been carried out previously, the majority have investigated single risk factors or used a relatively small sample size (Ismail et al., 2009, Johnson et al., 2016, Fisher-Owens et al., 2007, Harris et al., 2004). Similarly, there is a dearth of accurate and reliable data, at a national level, describing the pattern and amount of consumption of cariogenic food and drink during the preschool period (Gussy et al., 2016, Amezdroz et al., 2015).

Although dental caries is a diet mediated disease there is a recognised need for more studies investigating the complex relationship between diet and dental caries in the context of the child's living environment and especially focussing on modifiable risk factors (Wagner and Heinrich-Weltzien, 2017, Lee and Divaris, 2014, Casamassimo et al., 2014). It has been argued that for preschool children, in particular, the mechanisms for understanding how dietary intake affects ECC risk are poorly understood (Amezdroz et al., 2015, Johnson et al., 2016) and that most of the data relating to risk or protective factors have accumulated from cross sectional studies or from older child cohorts (Johnson et al., 2016, Gussy et al., 2016).

The global burden of dental disease and the urgent need to continue to untangle the complex nature of behavioural and physiological risk factors and socioeconomic determinants has been highlighted recently as crucial for reducing disease burden (Dye, 2017). The rapid growth of data science suggests that all health care disciplines need to be aware of these developments and integrate new data analysis techniques when addressing research questions (Wickham and Grolemund, 2016, Khoury et al., 2013). In epidemiology and population health, this impacts on academic training, knowledge integration, expertise and implementation of "Big Data" science, bioinformatics and other emerging technologies (Khoury et al., 2013). To advance population oral health the translation of knowledge discovery in these areas will still require an improvement of more upstream factors such as social determinants of health (Casamassimo et al., 2014).

Table 1.1. Selected International Child Cohort Studies.

Growing Up in Australia: The Longitudinal Study of Australian Children" (LSAC)	http://www.growingupinaustralia.gov.au/
Millennium Cohort Study (UK)	http://www.cls.ioe.ac.uk
Growing Up in New Zealand (GUiNZ)	http://www.growingup.co.nz/en.html
Growing Up in Ireland (GUI)	http://www.esri.ie/growing-up-in-ireland/
Growing Up in Scotland (GUS)	https://growingupinscotland.org.uk/

Large national longitudinal child surveys have attracted significant funding in recent years in many "developed countries" and provide a rich source of scientific data in multiple domains including health, wellbeing and nutrition. As many countries do not carry out regular national oral health surveys, which are resource intensive, combining oral health variables within these wide-domain child cohort surveys is a useful approach to interdisciplinary research within the framework of strong conceptual models of oral health (Divaris, 2016, Fisher-Owens et al., 2007, Kim Seow, 2012, Dye, 2017).

Using a data science approach to epidemiologic cohort analysis could yield valuable insight into multilevel interactions between variables and enable assessment of the contribution of each in causal models. Developing a predictive model for dental disease will require more longitudinal studies, high quality data measures and selection of the most suitable data-driven or investigator defined techniques (Divaris, 2016, Krebs-Smith et al., 2015, Harris et al., 2004).

The aim of this thesis was to utilise data analysis techniques, particularly decision tree methods, data mapping and association analysis, to explore multilevel interactions among key aspects of child and primary caregiver (PCG) psychosocial, behavioural and physical health affecting dental problems in a large national cohort of Irish preschool children. A specific emphasis was placed on investigating the role of diet as a common risk factor for weight status in the children and using meal association analysis to understand the pattern of cariogenic food intake. It was intended to explore the role of significant risk indicators in the context of previously published conceptual models for child oral health and development.

1.2. Dental problems in preschool children

1.2.1. Overview

Oral diseases were recently ranked in the top 10 leading causes of years lived with disability (YLDs) for the first time (Dye, 2017). The annual spending on oral care in the EU is approximately €79 billion and it is predicted that this may rise to €93 billion in 2020 (Patel, 2012). Recent evidence indicates that oral health disparities in young children may be widening (Casamassimo et al., 2014, World Health Organization, 2017b, Lee and Divaris, 2014). The lack of success

in attempting to integrate oral and general health policy and research strategies has, arguably, hampered efforts to prevent oral diseases (Lee et al., 2016, Sheiham, 2005, Jin et al., 2016). The lack of robust epidemiological data in the EU has also not assisted efforts to compare the current state of oral health across different countries and facilitate planning and resource allocation (Patel, 2012). In most countries, reliable and accurate data on the oral health-related problems and behaviours of children less than 5 years of age (preschoolers) is scarce. As with other chronic diseases, the early childhood dental problems are often neglected until later in the life-course when their sequelae are more expensive to treat and may have already caused significant social, psychological and emotional damage (Dye et al., 2010, Wagner and Heinrich-Weltzien, 2017).

Dental problems that result in a parent bringing a preschool child to the dentist include ECC, periodontal problems, malocclusion and trauma (Wagner and Heinrich-Weltzien, 2017). Dental pain is consistently associated with population levels of caries experience and this association is higher among those with reduced access to care (Slade, 2001). When people make the decision to visit the dentist for a problem they are much more likely to have a dental extraction than if they visit for a regular examination (Luzzi et al., 2013). ECC is the main reason for parental perceived child oral health problems and, in a public health system with limited access to dental care, the most common reason for a “problem visit” to a dentist (Declerck et al., 2008, Leroy et al., 2013, Luzzi et al., 2013). Children with toothache related to ECC are likely to need a dental procedure when they attend a dentist (Daher et al., 2015). Periodontal problems in infants are common and include soft tissue lesions, eruption cysts, gingivitis and periapical abscesses. A periapical abscess is, typically, a consequence of untreated ECC which results in an odontogenic infection and may require emergency treatment to prevent even more serious complications such as sepsis or cellulitis (Gussy et al., 2006, Wagner and Heinrich-Weltzien, 2017). Malocclusion, which is a developmental condition, may be associated with an increased risk of dental trauma. Developmental defects of dental tissues are associated with increased risk of ECC and tooth sensitivity.

ECC is a disease defined by the presence of one or more decayed, missing (due to caries), or filled tooth surfaces in any primary tooth in a child 71 months of age or younger (American Academy of Pediatric Dentistry, 2016). Any sign of smooth-surface caries in children less than 3 years of age, is indicative of severe early childhood caries (American Academy of Pediatric Dentistry, 2016).

Despite decades of reductions in dental caries levels ECC is a pandemic disease worldwide (Spencer, 2012, Bagramian et al., 2009, World Health Organization, 2017b). The prevalence of ECC is estimated to range from 12% to 70% depending on the population (Declerck et al., 2008, Gussy et al., 2016, Dye et al., 2015, Bourgeois and Llodra, 2014) with the highest rates occurring among those in immigrant and lower socioeconomic groups (Public Health England, 2013b, Dye et al., 2010).

There is a distinct lack of high quality data to describe the distribution of ECC in Ireland at a national level. A recent summary by O'Connell and Harding (2017) of the limited data from small sample size studies is shown in Table 1.2. The most notable outcome in those studies is that children with a lower socioeconomic status had higher mean d_{3vc}mft (decayed missing or filled primary teeth at visible, cavitated or non-cavitated, dentinal caries level) than those from a higher socioeconomic background and that ECC levels in those with disabilities was low when compared with that found in the general population.

In the last National Survey of Children's Dental Health (sampling occurred between 2001 and 2002), it was reported that 37% of 5-year-olds in fluoridated areas and 55% in non-fluoridated areas had experienced decay, i.e. they have one or more teeth that were decayed, filled or extracted as a result of decay (Whelton et al., 2006). Approximately 42% of children at 5 years of age had experienced dentinal caries and had a mean d_{3vc}mft of 1.3.

Most studies have shown that untreated ECC can lead to toothache and pain during the preschool period (Daher et al., 2015, Slade, 2001) and it is also the strongest predictor of dental caries in the permanent dentition (Gussy et al., 2006). It is well recognised that early preventive care is cost effective and the recommended age for a child's first dental visit is at no later than 12months of age (American Academy of Pediatric Dentistry, 2016). The children most in need of care are also least likely to receive it (Darmawikarta et al., 2014, Dye et al., 2010, Johnson et al., 2016). Unrestored dental caries is the main cause of dental pain in childhood (Boeira et al., 2012). Children can suffer with ECC and may experience pain, infection, altered weight status and difficulties with sleeping, eating and communicating (Boeira et al., 2012, World Health Organization, 2017b, American Academy of Pediatric Dentistry, 2016, Gussy et al., 2006, Bönecker et al., 2012).

Table 1.2 Summary of studies reporting caries in pre-school children in Ireland.

Authors	**Sample**	**% children with caries***
Holland, Houlihan and O'Mullane 1988	Pre-school children mean age 45 months (29 – 55 months)	38%
O'Connor 1996	Playschool children (mean age 4.1 yrs)	36% Fluoride group, 26% Non-Fluoride group
Tuohy 2000	3-year old healthy children	27.4% 41.5% Medical card 18% Non-Medical card
O'Connell *et al*. 2010	Children born small for gestational age (age 4-7 years)	53%
Happy teeth (baseline) 2013	Disadvantaged pre-school children n = 233	18.5%
Sagheri, McLoughlin and Nunn 2013	Children with disabilities >4 years old, n=337	Age 3 = 0% Age 4 = 18.2% Age 5 = 24.6%
Stapleton 2015	Children with disabilities age 0-6 years, n = 178	Age 3 = 0%

Source: Adapted from O'Connell and Harding (2017).

* Caries recorded at the dentine level, with or without cavitation (d_{3vc}mft)

The last National Survey of Children's Dental Health in Ireland indicated that approximately 83% of dentinal caries present in 5- year olds was untreated (Whelton et al., 2006). Consequently, it is more usual for most preschool children with extensive ECC to attend for emergency care which often results in dental extractions and sometimes requires hospitalisation for general anaesthesia (Smith et al., 2014, Slack-Smith et al., 2009, O'Connell and Harding, 2017). One of the key recommendations of the report from the 'WHO expert consultation on public health intervention against early childhood caries' was to include the three-year-old age group as one of the index ages recommended for population surveys in the next edition of WHO's Oral health surveys: basic methods (World Health Organization, 2017b).

There is increasing recognition of the importance of oral health for the general health and development of infants and young children (American Academy of Pediatric Dentistry, 2016, Tinanoff and Reisine, 2009, Petersen and Kwan, 2011). Novel multidisciplinary longitudinal research projects such as the

VicGeneration (VicGen) study of an Australian oral health birth cohort are much needed to develop a solid understanding of risk and protective factors for the development of ECC (Gussy et al., 2016, Johnson et al., 2016). Child oral health research should be integrated within other health and development research frameworks to fully elucidate any common risk factors and causal pathways (Casamassimo et al., 2014, Divaris, 2016, Lee and Divaris, 2014).

1.2.2. Natural history of ECC

It is important to understand the natural history of dental caries so that intervention and prevention strategies can be implemented (Leong et al., 2013, Gussy et al., 2016). Although ECC has multifactorial origins it essentially occurs through the metabolism of fermentable dietary carbohydrates at the plaque-biofilm interface resulting in localised demineralisation and destruction of hard dental tissues over time (Selwitz et al., 2007, Bradshaw and Lynch, 2013, Gussy et al., 2006). Dental caries has been described as “an infectious disease modified by diet” (Gussy et al., 2006) although others have argued that “the pivotal role of sugars in causing caries differs markedly from the widely held erroneous notion that caries is a multifactorial infectious transmissible disease” (Sheiham and James, 2015). However, this view is somewhat controversial, and it is generally accepted that while dietary sugar is uniquely cariogenic, understanding the factors that influence the oral microbiome is necessary to explain the pathogenesis of ECC and the links between oral and general systemic health (Gomez et al., 2017, Do et al., 2013). Fundamentally, the causative model of dental caries includes the interaction of the host (susceptible tooth surface), environment (dietary carbohydrate) and agent (oral biofilm) over time (Marsh, 2006, Gussy et al., 2006, Selwitz et al., 2007).

It has been suggested that the influence of modifiable risk and protective factors on early childhood caries, and the optimal age at which preventive measures should be taken has not been well established (Gussy et al., 2016). Even in the pre-dentate stage the bacteria present play a significant role in early caries experience with colonisation mediated by feeding practices and oral health related behaviours (Leong et al., 2013). The VicGen study in Australia has tracked the natural history of dental caries in children from birth and found very little dental caries activity in the first 18 months (Johnson et al., 2016). However, they reported a dynamic period between 18-40 months with rapidly developing new lesions and regression of lesions from non-cavitated lesions to sound teeth suggesting this may be a key period for introduction of both risk factors (such

as cariogenic food and drink) and protective factors (such as tooth brushing with fluoridated toothpaste) (Gussy et al., 2016). These findings and others from studies of early life feeding behaviours (Chaffee et al., 2015) emphasise the importance of preventive and behavioural interventions at this critical period of development to change the life-course trajectory of dental problems.

1.2.3. Risk factors and interdisciplinary approaches

1.2.3.1. Terminology

Over 100 risk factors for ECC have been identified and comprehensive reviews reported in the literature (Leong et al., 2013, Harris et al., 2004, Gussy et al., 2006). Given the relatively young science of epidemiology a rather casual approach to the use of terminology appears to pervade most publications (Krieger, 2008, Krieger, 2012). For example, although the term risk factor is generally accepted to indicate an exposure that is related to an outcome (statistically) it is not clear in the literature whether a risk factor is a causal link or an association (Burt, 2001). Terms related to risk such as risk factor, risk indicator, determinant modifiable risk factor and risk marker have all been used in the literature without being well defined. In oral health research a risk factor has been defined as “an environmental, behavioural, or biologic factor confirmed by temporal sequence, usually in longitudinal studies, which if present directly increases the probability of a disease occurring, and if absent or removed reduces the probability” (Beck, 1998). Burt (2001) described a risk indicator as “a probable, or putative, risk factor, but the cross-sectional data upon which it is based is weaker than the results of longitudinal studies”. In dental research It has been argued that a broader view of risk would account for the social determinants of health and population health (Burt, 2005). Rather than solely focus on “downstream” biological or behavioural factors a shift towards investigating the “upstream” or distal determinants of dental problems in the population is required to address the social gradient (Newton and Bower, 2005). Population level parameters are often incorrectly applied to individuals and this concept has been defined as the “privatisation of risk”. This is misleading as risk can be determined for populations or subgroups in longitudinal studies, but risk factors are not the “cause” of individual cases of disease (Rockhill, 2001, Divaris, 2016). This explains why risk factors for ECC are poor predictors of individual cases of disease occurrence but continue to be strongly associated with the prevalence or incidence of ECC (Divaris, 2016).

The Rose hypothesis suggested that as long as disease risk is widespread, then adopting measures that decrease risk for everyone is more effective in reducing disease burden than taking a 'high risk' approach, in which measures are targeted only to those individuals with a substantially increased risk for disease (Rose, 1992). A combination of both the "high risk strategy" and population oral health approaches has been adopted by many countries, but some have argued that measures for identifying the high-risk subgroup in a population are neither sufficiently accurate or reliable (Tickle and Milsom, 2008, Hausen, 1997). However, to advance the potential of both approaches in ECC risk assessment key steps are required which include large longitudinal cohorts with high-quality clinical examinations, valid preclinical disease markers and rigorous predictive modelling (Divaris, 2016).

1.2.3.2. Food and drink factors and oral health

A clearer understanding of how modifiable and non-modifiable factors affect oral health problems is necessary so that appropriate intervention policy and strategies can be developed to maximum effect. (Leong et al., 2013, Chi et al., 2017, Chankanka et al., 2015). Much of the early research in ECC focused only on the contribution of cariogenic food and drink to the aetiology and prevention of lesions (Gussy et al., 2006, Leong et al., 2013). While the availability of refined carbohydrate is required for ECC to occur there is now an increased understanding of the interactions between other biological, behavioural and psycho-social risk factors. The potential cariogenicity of frequently provided human milk or infant formula or cow's milk given in a bottle is controversial (Leong et al., 2013, Gussy et al., 2006, Harris et al., 2004). The current evidence indicates that breastfeeding up to 1 year of age is associated with a reduction in dental caries risk while there may be an increased risk of ECC with on-demand breast feeding or nocturnal feeding and sleeping with the breast in the mouth (Tham et al., 2015). While the addition of sugars in bottle feeding is strongly associated with ECC the lack of good quality studies and control of confounding factors suggests that further research is required to inform feeding guidelines (World Health Organization, 2017b, Gussy et al., 2016).

1.2.3.3. Interdisciplinary factors

There have been many calls for researchers to focus more on interdisciplinary research, but few studies have done so particularly at the preschool age (Lee et al., 2016, Casamassimo et al., 2014, Fisher-Owens et al., 2007, Divaris, 2016). Few longitudinal studies in preschool children have adopted a

multidisciplinary approach although the VicGen study is an example of one which has measured maternal and child health, oral health related behaviours, behaviours, attitudes, knowledge, socio-demographics and diet, carried out oral assessments and taken saliva samples to determine the microbiome (Johnson et al., 2016).

The bidirectional relationship between oral health problems and child health and development is complicated by a variety of sociodemographic influences (Hooley et al., 2012a, Sheiham, 2006). In addition, the primary caregiver (PCG) is the gate-keeper in providing and promoting general and oral health care for the developing child; therefore, PCG health and wellbeing is intricately linked to child health and, ultimately, defined by similar social determinants (Moimaz et al., 2014). Dental research has expanded in recent years, recognising that psychosocial, behavioural and environmental factors significantly impact oral health outcomes (Fisher-Owens et al., 2007, Newton and Bower, 2005). A number of studies have reported relationships, between PCG psychological distress, child socio-emotional behaviour or infant temperament and child oral health outcomes. (Tang et al., 2005, Menon et al., 2013, Quinonez et al., 2001a, Spitz et al., 2006, Aminabadi et al., 2014). Parental stress and depression may impact on the caregivers' ability to impart preventive oral health measures at vulnerable developmental stages (Tang et al., 2005, LaValle et al., 2000) and are often related to aspects of infant temperament and child socio-emotional behaviour (Spitz et al., 2006, Renzaho and Silva-Sanigorski, 2013, Mäntymaa et al., 2006). Depressive symptoms in mothers may lead to inconsistent parenting and unhealthy feeding habits (Kim Seow, 2012). A positive child temperament appears to be protective against early childhood caries (ECC) while a negative temperament and poor feeding practices are both equally strongly associated with ECC (Aminabadi et al., 2014). Interestingly, while reporting of subject ethnicity has been widespread, very few studies have included ethnic background as a predictor or independant variable in subsequent analyses (Harris et al., 2004), although in recent years ethnic disparities in oral health has received some attention (Fisher-Owens et al., 2013, Riggs et al., 2015). Few studies have examined the role of psychosocial and behavioural factors on oral health in large population studies (Hooley et al., 2012b). Furthermore, most research to date, focussed on early childhood dental problems and the health and psychosocial attributes of the child and PCG, has concentrated on the effect of a single variable using relatively small sample sizes (LaValle et al., 2000, Hooley et al., 2012b, Abreu et al., 2015).

1.2.3.4. Prevention of ECC and the common risk factor

In the early preschool stage of development, the infant is totally reliant on the mother or PCG in determining the dietary intake, oral hygiene behaviours and health care needs of the child (Wagner and Heinrich-Weltzien, 2017). Even in the first year of life the infant vertical transmission of oral cariogenic bacteria can occur in the pre-dentate stage and play an important role in development of dental caries while protective maternal interventions such as healthy eating habits and regular toothbrushing with fluoridated toothpaste can reduce the risk of ECC (Leong et al., 2013). Early intervention is crucial to establish good oral hygiene and dietary practices, especially as studies have reported that consumption of "unhealthy" or discretionary choice foods and beverage is higher than the recommended guidelines in 20-49% of children even before 3 years of age (Amezdroz et al., 2015, Crowe et al., 2017). While dental caries can be arrested and, under the appropriate conditions, even reversed in its early stages, it is often allowed develop further by inappropriate feeding practices and significant infection with cariogenic bacteria (mutans streptococci) (Gussy et al., 2006, Johnson et al., 2016, Chaffee et al., 2015). Other risk factors for ECC include genetics, saliva, enamel hypoplasia, oral hygiene behaviour and family environment (Gussy et al., 2006, Kim Seow, 2012). While dental plaque and poor toothbrushing habits are strong risk factors for ECC (Meurman and Pienihäkkinen, 2011, Gussy et al., 2006, Harris et al., 2004) baseline experience of dental caries is still the single most important predictor of future caries experience (Gussy et al., 2006, Harris et al., 2004).

Despite strong evidence showing that the benefits of preventing dental caries greatly outweigh the treatment costs there is a clear lack of emphasis on prevention policies and strategies throughout the EU (Patel, 2012). Adopting a common risk factor approach to prevention strategies and policy approaches has been widely recommended (Wagner and Heinrich-Weltzien, 2017, Public Health England, 2015a, Sheiham and Watt, 2000, Watt and Sheiham, 2012, Petersen and Kwan, 2011, World Health Organization, 2000, World Health Organization, 2017b). Whether the population approach, high risk approach or a combination of both are adopted, it is clear that all prevention strategies require a comprehensive understanding of both protective and risk factors for ECC (World Health Organization, 2017b, Gussy et al., 2006, Johnson et al., 2016, Chaffee et al., 2015, Leong et al., 2013). Primary prevention strategies currently include oral hygiene instruction, dietary management and application

of fluoride. Clearly, current strategies to prevent ECC have not been fully successful (Harris et al., 2004) and adopting the common risk factor approach and integrating interdisciplinary aspects of managing risk factors within primary health care is needed (World Health Organization, 2017b).

Many of the steps to progress the development of valid and efficient predictive modelling of ECC are already well elucidated by other researchers (Divaris, 2016, Casamassimo et al., 2014). A clear understanding of the differences between "ECC risk assessment at the population level and 'precision dentistry' at the person level and personal" is required to advance these steps (Divaris, 2016).

1.3. Conceptual models of child oral health

1.3.1. Overview

The life-course approach to epidemiology has been defined as: "The study of long term effects on later health or disease risk of physical or social exposures during gestation, childhood, adolescence, young adulthood and later adult life" (Kuh et al., 2003). Kuh further stated that: "The aim is to elucidate biological, behavioural, and psychosocial processes that operate across an individual's life course, or across generations, to influence the development of disease risk". While investigating a birth cohort is one part of this approach lifecourse, epidemiology is more than a study of accumulated risk or protective factors in a longitudinal study and requires a theoretical model and a study design. A better understanding of the relevant life course factors and oral health problems may be achieved by merging data from different studies and disciplines (Khoury et al., 2013, Abreu et al., 2015). Given the public health need for targeted interventions aimed at reducing the prevalence of dental disease, particularly ECC, it is vital to develop the knowledge and skills-base to understand and identify protective and risk factors using both conceptual models and interdisciplinary databases that have even limited measures of oral health. This is especially true in countries, including Ireland, where national oral health surveys have not been undertaken in more than a decade and have never included preschool children.

The development of strong conceptual models for children's oral health in recent years has been a feature of dental epidemiology in researching multilevel factors that can have an impact at the early life stage and throughout the life course (Nicolau et al., 2007, Ben-Shlomo and Kuh, 2002, Fisher-Owens et al., 2007). The role of social and behavioural factors in developing ECC has become more evident in the last decade (Ismail et al., 2009, Ismail, 2003, Finlayson et al., 2007, Reisine et al., 1994). Similarly, research in child health and development has focussed on broader bioecological models with the child at the centre of a complex, multi-layered, interconnected, environmental system which influences child development. "The bioecological model, together with its corresponding research designs, is an evolving theoretical system for the scientific study of human development over time" (Bronfenbrenner and Morris, 2007). Upstream factors such as family functioning (Castilho et al., 2013, Duijster et al., 2013b, Duijster et al., 2013a), social class (Gibson and Williams, 1999, Boyce et al., 2010), parental influences (Renzaho and Silva-Sanigorski, 2013, Hooley et al., 2012b) and psychosocial factors (Quinonez et al., 2001b, Reisine et al., 1994, Finlayson et al., 2007) have all been studied in relation to child dental health. Recent commentaries (Lee and Divaris, 2014, Sniehotta et al., 2017, Divaris, 2016) have highlighted the need for an integrated multidisciplinary approach to population-based health/oral health research rather than individually examining the effects of upstream or downstream risk factors (Krieger, 2008).

A more modern and functional definition of oral health by FDI World Dental Federation includes: "Further attributes include that it is a fundamental component of health and physical and mental wellbeing. It exists along a continuum influenced by the values and attitudes of individuals and communities; it reflects the physiologic, social, and psychological attributes that are essential to quality of life" (Glick et al., 2017). Lee et al (2016) further noted that this new definition for oral health created a challenge for the research community to "develop and evaluate a consensus set of measures of the domains of oral health that will be adaptable to the questions within target clinical disciplines". Influenced by behavioural and social science research multiple conceptual models of oral health have been developed to incorporate a wider framework of domains which expand beyond the individual, to the family and community level (Fisher-Owens et al., 2007, Kim Seow, 2012, Casamassimo et al., 2014, Lee and Divaris, 2014). Most recently, a causal model was constructed (Figure 1.1), adapted from the Fisher-Owens model,

based on repeated collection of comprehensive survey data, including oral health assessments, of the VicGen birth cohort over an eight year period (Johnson et al., 2016). This type of causal pathway should allow for a predictive modelling approach to determine the contribution of the relevant risk factors on child oral health.

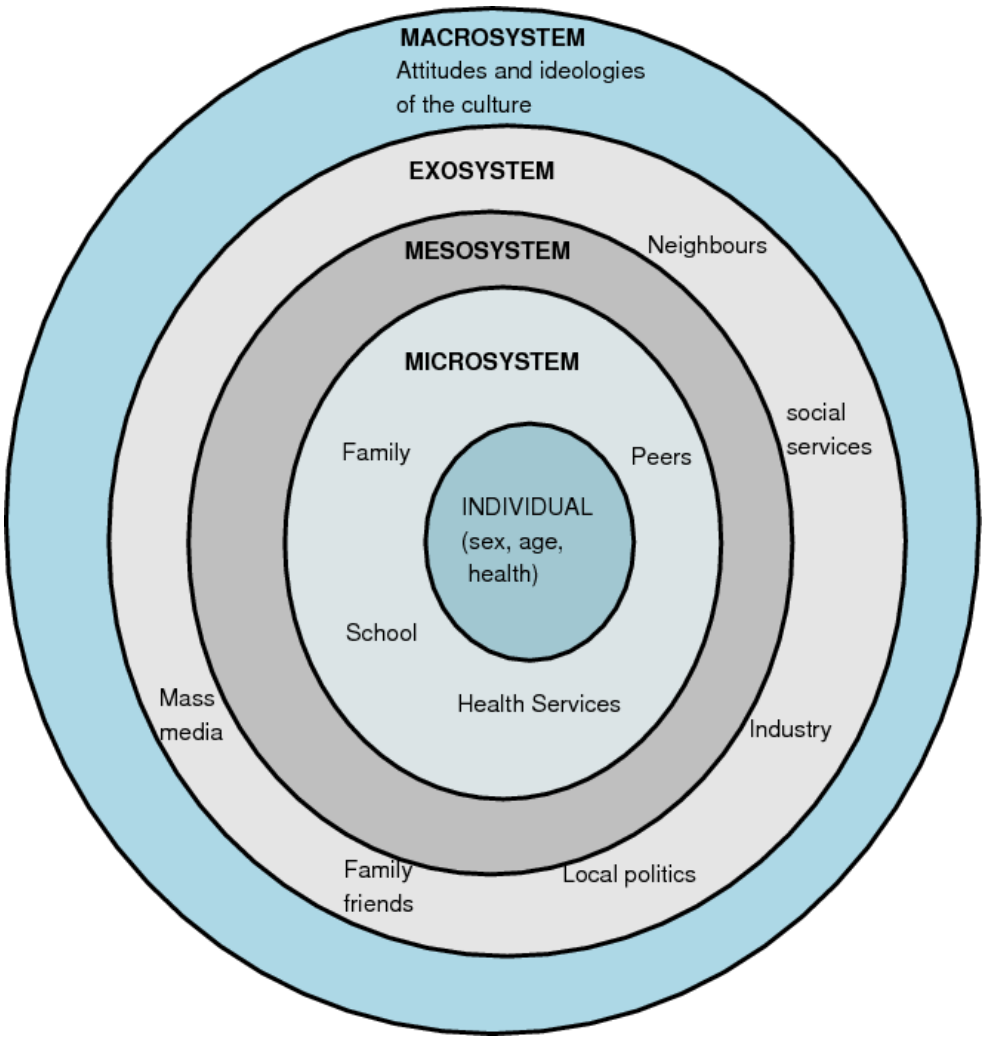


Figure 1.1 The VicGeneration Causal Model of child, family and community influences on oral health outcomes for children. Amended from Johnson et al. (2016).

There is a pressing need to optimise the use of all resources, both funding and data, for epidemiological studies. Khoury *et al* (2013) recently emphasised the

importance of “transforming epidemiology for 21st century medicine and public health” and highlighted the need to focus on interdisciplinary collaboration, greater access to data to ensure reproducibility, expand cohort studies across the lifespan and integrate “big data” science into epidemiological practice. The report also recommended the promotion of cross-study best practices for managing datasets and development of novel analytic strategies.

1.3.2. National child surveys

There is, currently, no national data in Ireland on early childhood dental health and the last national dental survey, which only included children five years of age and older, was carried out 15 years ago (Whelton et al., 2006). Few countries have measured or surveyed dental related measures or outcomes at a national level for preschool children. However, many countries have developed and funded national cohort surveys and most of these have included some oral health related measures. Some (e.g. LSAC) have also carried out subgroup clinical examinations or nested studies within the cohorts (Wake et al., 2014). Some of these nested studies include a wide range of physical health measures and biosamples which can be stored for biochemical, genetic and epigenetic analysis. As mentioned earlier, preschool children have not been previously included in national oral health surveys in Ireland. In recent years one of the best opportunities for capturing nationally representative data on oral health related measures such as diet, oral hygiene and dental problems in preschool children derived from cohort surveys such as GUI (Quail et al., 2011, Murray et al., 2013) (http://www.esri.ie/growing-up-in-ireland) and the National Preschool Nutrition Survey (NPNS) (Irish Universities Nutrition Alliance, 2012) (http://www.iuna.net).

1.3.3. Growing Up in Ireland

“Ensuring the good health and well-being of all young children in Ireland” is one of the central goals of the *Report of The Expert Advisory Group on The Early Years Strategy* (Department of Children and Youth Affairs, 2013). The Department for Children and Youth Affairs (DCYA) has invested €35 million in the Growing Up in Ireland (GUI) study which provides comprehensive data on aspects of child development that will provide a statistical basis to inform policy and “ensure a better future for Ireland’s children” (Department of Children and Youth Affairs, 2014). The GUI study is the largest, most significant, child

research project ever carried out in the state. Key outcome domains include physical health and development and socioemotional well-being (Williams et al., 2013). GUI is a national longitudinal study following the progress of two groups of children: the Child Cohort includes 8,500 nine-year-olds; and the Infant Cohort includes 11,000 nine-month-olds born in the Irish Republic who participated in the study at 9 months, 3 years and 5 years of age (Figure 1.2) (http://www.growingup.ie). The study is being carried out by a group of researchers led by the Economic and Social Research Institute (ESRI) and Trinity College Dublin (TCD). Funding is provided by the Department of Children and Youth Affairs in association with the Department of Social Protection and the Central Statistics Office. GUI began in 2006 and currently the members of the infant cohort are 10 years old while those of the child cohort are approximately 20 years old.

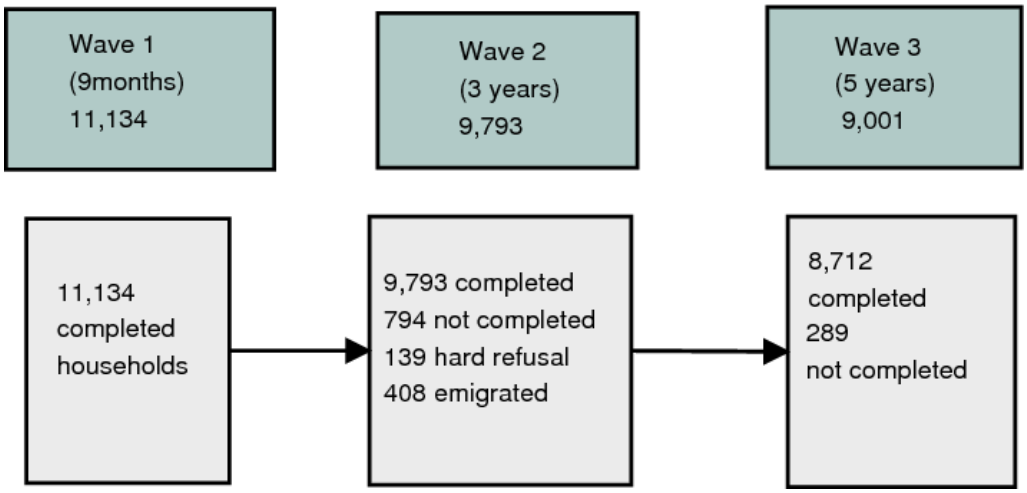


Figure 1.2 Infant Cohort Longitudinal Sample, Growing Up in Ireland Survey. Note: 11,134 questionnaires issued at Wave 2; 10,587 issued at Wave 3. Questionnaires not completed at Wave 2= 794; not completed at Wave 2 but included in Wave 3=289.

GUI is the most comprehensive study of children ever undertaken in Ireland and the primary objectives include a description of child development over time and investigating how different experiences in early childhood may affect outcomes at an older age. A key focus is to provide an appropriate evidence base to facilitate development of child and family policies (Williams et al., 2013).

The study emphasises the bioecological approach to understanding child development which places the child at the centre surrounded by multiple influencing layers that include the child's characteristics, family, neighbours and community services that occur over time (Figure 1.3) (Bronfenbrenner and Morris, 2007).

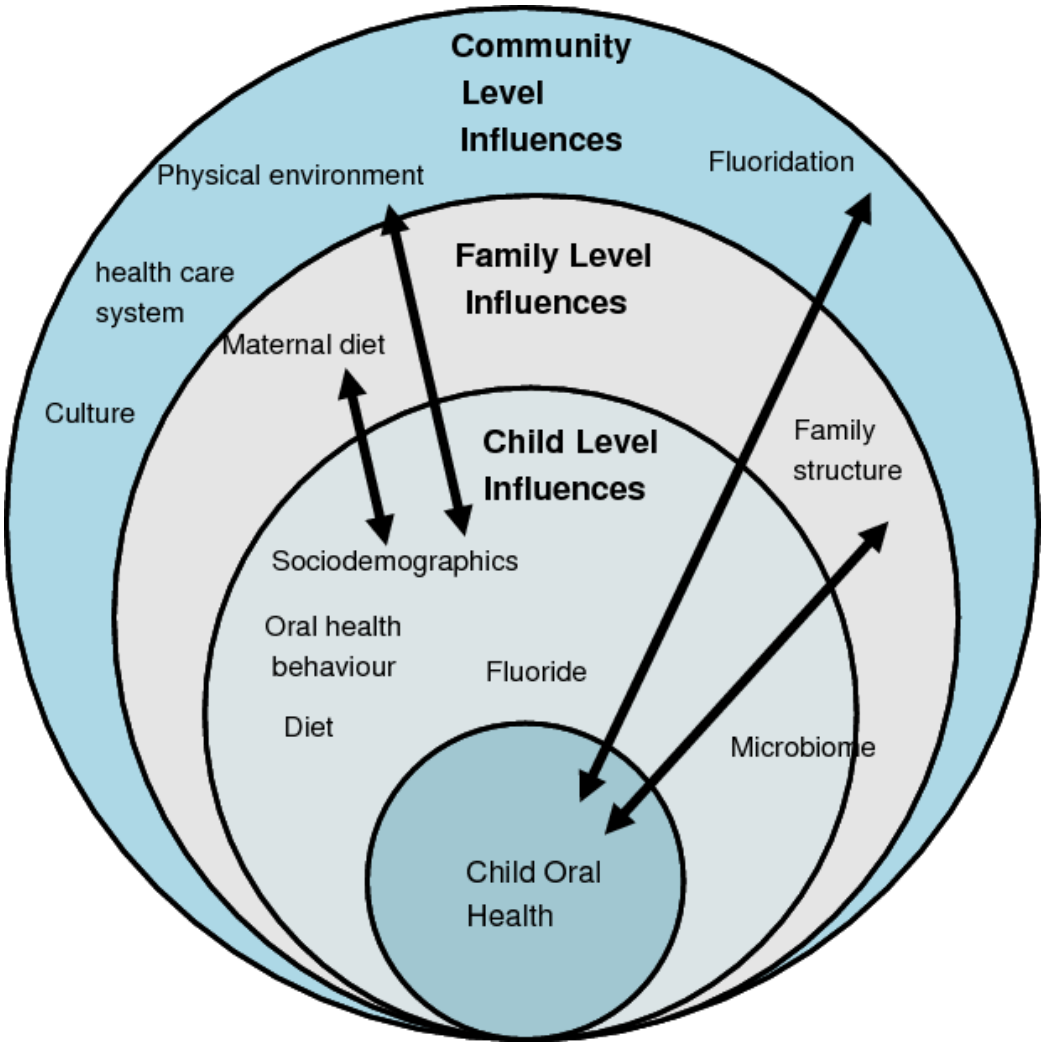


Figure 1.3 Bronfenbrenner's Bioecological Model of Human development. Amended from Garbarino (1982).

This model also appears to have influenced the conceptual oral health models developed more recently by Fisher-Owens (2007) and others (Casamassimo et al., 2014, Johnson et al., 2016, Kim Seow, 2012).

GUI has concentrated on three main aspects of child development:

- Physical health and development
- Social, emotional and behavioural well-being
- Educational achievement

From the perspective of examining oral health issues, a key strength of this nationally representative study is the multidisciplinary nature of the data collected. The dataset contains a comprehensive range of measures related to both parents and children concerning general, psycho-social and behavioural health. The themes of the health domain included: Pregnancy / pre-natal care; Child's birth; Child's health/healthcare utilisation; Child's nutrition/diet/ breastfeeding; Child's physical activity levels/exercise; Child's physical development; Physical measures and Parental health and lifestyle.

In the infant cohort, questions related to the child's oral health included dental problem visits, oral behavioural habits including frequency of tooth brushing and how often the child sucked a soother or finger/thumb at 9 months of age and accessing dental care at 5 years of age. Information relating to diet was also collected using a food frequency questionnaire. Other measures included global health rating, detailed data on breastfeeding, illness/hospital attendance, socio-demographics and psycho-social variables including temperament and behaviour. The only physical measures taken were the length and weight of the infant. There are multiple publications arising from this cohort including investigations of infant feeding (Castro et al., 2015, Castro et al., 2014) and obesity (Layte et al., 2014a).

1.3.4. National Preschool Nutrition Survey (NPNS)

The first NPNS study was conducted by the Irish Universities Nutrition Alliance (IUNA) from October 2010 to September 2011. As with the scarcity of oral health data for this age group there has been a similar lack of detailed estimates of dietary intake for preschool children prior to this survey. Similar to problems with questionnaire-based reporting of oral health related measures, there are particular difficulties in estimating the pattern of consumption of food and drinks at this age as the information is usually reported by a parent or guardian.

IUNA has completed a number of national nutritional surveys among various other age groups including children (2003), teenagers (2005) and adults (2011) (http://www.iuna.net). Analysis of these datasets has provided valuable

information on nutrient intake, food safety, informed food policy and development of food based dietary guidelines. National dietary surveys among preschool children are necessary for understanding and describing food consumption patterns and nutrient intake and allow identification of consumption that does not meet recommended dietary goals (Gibney and Wolmarans, 2004). The NPNS also collected anthropometric measures and information related to child health and lifestyle.

1.4. Diet, obesity and dental health- the common risk factor approach

Oral health is inextricably linked with general health, especially with self-reported health status (Lee et al., 2013, Sheiham, 2005, US Department of Health and Human Services, 2000, Fisher-Owens et al., 2007). Poor nutritional behaviours established in preschoolers are associated with an increased risk of obesity and dental caries that can affect the individual throughout the life course (World Health Organization, 2002, Kuh et al., 2003).

Obesity may be defined as an excess of body fat (Flegal and Ogden, 2011). Body mass index (BMI) has only relatively recently been used for assessing child weight/fatness and is more complicated than its use in adults as it is age and gender specific with no universal agreement on growth reference charts or cut off criteria to use for classifying obesity and overweight (Flegal and Ogden, 2011, Rolland-Cachera, 2011). However, most studies tend to classify child BMI under age 5 years using the International Obesity Task Force (IOTF) and WHO cut-offs (Cole et al., 2000, Onis, 2006, Cole et al., 2007). In 2013, 42 million preschool children, world-wide, were classified as obese or overweight (de Onis et al., 2010, de Onis and Lobstein, 2010). Although the prevalence of overweight and obesity in Irish children appears to be stabilising (Keane et al., 2014), using the limited national data for preschoolers and the IOTF cut-offs, the prevalence of obesity and overweight in 2-4 year old children was 15% and 3% respectively. In the same sample from NPNS, categorised using the UK/WHO growth charts, obesity and overweight was estimated as 16% and 7%, respectively (Irish Universities Nutrition Alliance, 2012). In Europe, 12-15% of preschool children were classified as overweight or obese based on IOTF criteria (Ahrens et al., 2014, Walton, 2012). Similarly, epidemiological data on the dental health of preschool children in Ireland is minimal and this age cohort

(less than 5 years) have not previously been included in national dental surveys. Given the relatively sparse studies of obesity in preschool children globally it has been suggested that population surveys of this age group are urgently required to both assess prevalence of overweight and obesity and also attempt to identify potential patterns and susceptibility of subgroups (Wake et al., 2007).

The primary caregiver (PCG) plays a key role in facilitating prevention of both obesity and ECC through feeding patterns and other behaviours in the preschool period (Wake et al., 2008, Pine, 2013, Gussy et al., 2006, Dye et al., 2004, Chaffee et al., 2015). Obesity and dental caries share some common risk factors such as food choice, patterns of dietary intake, poor quality diet and socioeconomic factors such as education level of PCG and household income (Marshall et al., 2007b, Kantovitz et al., 2006, Sheiham and Watt, 2000, Watt and Sheiham, 2012). Dental caries and weight status are diet-related, and patterns of nutritional behaviours are established in early life (Wake et al., 2007, Chaffee et al., 2015). However, very few epidemiological studies have reported anthropometric and dental health measures of preschool children, despite the fact that both early childhood obesity and dental caries are strong predictors of these respective conditions throughout the life course (Dye et al., 2015, Wake et al., 2007).

Many of the cross-sectional and longitudinal studies that explored the association between diet, dental caries and obesity have found conflicting results (Alves et al., 2013, Costacurta et al., 2014, Hong et al., 2008, Liang et al., 2016, Goodson et al., 2013, Qadri et al., 2015). Although some studies have shown a positive relationship between BMI and dental caries, others suggest that they are correlated weakly and that different predictors may be associated with dental caries at both high and low BMI levels (Hooley et al., 2012a, Kantovitz et al., 2006, 2015a, Arcella et al., 2002, Marshall et al., 2007b). This should not be too surprising given the potential for confounding and the complex aetiology of obesity, in particular. Although there is ample evidence that dental caries will not develop without dietary sugar (Anderson et al., 2009), numerous epidemiological studies have failed to find an association, or found a low association between frequency of intake of carbohydrates and obesity (Te Morenga et al., 2013, Hayden et al., 2012, Goodson et al., 2013, Stanhope, 2015). While evidence for the association of dental caries and sugar consumption is derived from a wide variety of studies it is tacitly accepted that a high sugar diet is a “necessary cause” of dental caries (Tinanoff et al., 2002,

Rothman et al., 2008, Goodson et al., 2013). Recently, there appears to be an increasingly controversial focus in identifying sugar as the primary cause of obesity and a significant contributing factor to many non-communicable diseases in the so-called “developed” world (Lustig et al., 2012, Stanhope, 2015, Pereira, 2006). It is generally accepted that excess caloric intake from any macronutrient can lead to conditions such as overweight and obesity. However, the current controversy relates to the argument that dietary sugar is uniquely toxic and addictive and an independant cause of obesity (Lustig et al., 2012, Bray and Popkin, 2014) while others have argued that obesity and overweight continue to increase in countries where, at the same time, there have been substantial declines in sugar consumption (Barclay and Brand-Miller, 2011, Brand-Miller and Barclay, 2017).

Interestingly, there have been few reports on the oral health status of underweight children and it is uncertain whether they have a low caries risk (Tramini et al., 2009). It has been suggested that dental caries risk factors for underweight children needs to be assessed independently and not included in the “normal” weight category (Hooley et al., 2012a). There is some evidence to suggest that dental caries may be associated with children who are underweight and suffer with slow growth due to pain on mastication (Wagner and Heinrich-Weltzien, 2017, Hooley et al., 2012c, Sheiham, 2006, Tramini et al., 2009). This may inhibit growth hormone release via pain inducing altered sleep patterns and increased glucocorticosteroid secretion, (Vania et al., 2011) the effects of which may be further heightened by undernutrition due to an inability to chew more healthy foods (Gussy et al., 2006).

As reduced socioeconomic status and a high intake of dietary sugars are risk factors for both obesity and dental caries there is a strong conceptual basis for adopting interventions that address these common risk factors (2015a, Goodson et al., 2013, Qadri et al., 2015). Other common risk factors include psychosocial, lifestyle and family/school based patterns (Crowe et al., 2016, Watt and Sheiham, 2012, Sheiham and Watt, 2000). While obesity and dental caries are strongly associated with lower socioeconomic status the causal pathways have still not been fully elucidated. Furthermore, there are many other risk factors for both conditions that are not shared (Public Health England, 2015a). The evidence that obesity and dental caries may be more likely to occur in the same populations was found to be of “low-quality” according to a recent summary report by Public Health England (2015a) which was largely based on two systematic reviews (Hayden et al., 2012, Hooley et al., 2012a). As these

were mainly compiled using cross-sectional data it was not possible to conclude if “an individual who is overweight or obese is at higher risk of dental caries or vice-versa” (Public Health England, 2015a).

Given the associations that exist between oral health and general health (Sheiham, 2005, Pine, 2013) there is growing interest in utilising a common risk factor approach to investigate the multidimensional causes of dental and weight status problems, particularly in preschool children (Vania et al., 2011, Hooley et al., 2012a, Hooley et al., 2012c, Hayden et al., 2012, Costacurta et al., 2014, 2015a). The main focus of what is described as the “common risk factor” approach is to control a small number of overlapping or shared risk factors between chronic diseases by promoting general health and thus exert a greater, more efficient, impact than disease specific approaches (Sheiham and Watt, 2000).

The Global Burden of Disease (GBD) study has indicated that the transition of less developed countries to more economically developed, will also result in an increase in chronic disease while infectious disease declines. The only feasible long-term strategy for global oral health “must focus on health promotion and disease prevention, through controlling the modifiable common risk factors” (Jin et al., 2016). It is vitally important to identify common risk factors to accelerate future oral health research targeted at prevention of ECC and childhood obesity (Chi et al., 2017, Divaris, 2016, Watt and Sheiham, 2012). Using a common risk factor approach to address both conditions is sensible as long as the evidence for the role of sugar in causing obesity is convincing (Watt and Sheiham, 2012). As highlighted by Dye (2017), the “epidemiological transition” will impact on oral health globally as oral diseases share common risk factors such as an unhealthy diet with other non-communicable diseases (NCDs).

1.5. Cariogenic food and drink

1.5.1. Overview

As mentioned previously, consumption of cariogenic food and drink (CF) contributes to ECC (Hujoel and Lingstrom, 2017, Chankanka et al., 2015, Newens and Walton, 2016, World Health Organization, 2017b). This section will examine research that focuses on patterns of consumption of CF (Hujoel and Lingstrom, 2017, Chankanka et al., 2015, Newens and Walton, 2016,

World Health Organization, 2017b, Kassebaum et al., 2015). The carious process occurs through the metabolism of fermentable dietary carbohydrates at the plaque-biofilm interface resulting in localised demineralisation and destruction of hard dental tissues over time (Selwitz et al., 2007, Bradshaw and Lynch, 2013). While evidence describing the impact of nutrition intervention during crucial growth periods is well established (Bhutta et al., 2013), it appears to be surprisingly little known that the early stages of dental caries can be halted and potentially reversed (Selwitz et al., Lingström, 2009, Moynihan and Petersen, 2004) and that food intake patterns play a key role in dictating the progression or reversal of the disease (Moynihan and Kelly, 2014, Gussy et al., 2016, Chaffee et al., 2015, Tinanoff et al., 2002).

In a sufficient-cause model (Rothman et al., 2008) dental caries cannot occur without consumption of CF, especially sugar, and it is generally accepted that this is a 'necessary cause' of dental caries (Tinanoff et al., 2002, Sheiham and James, 2015). However, despite improvements in understanding many of the proximal and distal determinants of ECC, (Selwitz et al., 2007, Divaris, 2016, Fisher-Owens et al., 2007) an observation made more than 50 years ago remains valid today: "no one yet knows the precise nature of dental caries and all of the ramifications of the consumption of carbohydrates that influence caries" (Peterson, 1963). While some studies have explored the relationship between the pattern of consumption of CF and dental caries in young children (Chaffee et al., 2015, Dye et al., 2004, Johansson et al., 2010, Marshall et al., 2005, Llena and Forner, 2008, Hooley et al., 2012b) detailed understanding of the frequency, amount and pattern of intake of CF in preschool children is still limited (Marshall et al., 2005, Dye et al., 2004, Chankanka et al., 2015, Moynihan and Kelly, 2014). While a wide range of risk indicators have been found to be associated with ECC such as a cariogenic diet, poor oral hygiene, low fluoride exposure, inappropriate feeding methods, poor parental education and household income, very few studies have looked at risk factors using the appropriate longitudinal study design and validated dietary instruments (Hooley et al., 2012b, Harris et al., 2004, Selwitz et al., 2007).

Given the changing and dynamic nature of dietary habits in children it is important to understand the typical intake of CF as snacks or part of main meals (Llena and Forner, 2008, Johansson et al., 2010). A diet containing a greater amount of CF before age 12 months is associated with an increased incidence of dental caries by preschool age (Chaffee et al., 2015). Parental attitudes and eating practices strongly influence the eating behaviours developed in

childhood (Fisher et al., 2015). While infants are born with an innate preference for sweet taste (Stephen et al., 2012) there is strong evidence to suggest that learning to develop a liking for "healthy" foods is modifiable early in life (Mennella, 2014). The early introduction of CF can foster a long-term pattern of preference for the consumption of CF (Marshall et al., 2007a) and also increase the virulence and levels of cariogenic bacteria such as mutans streptococci and lactobacili (Marsh, 2006, Chaffee et al., 2015). However, this is also a crucial period where the deleterious effects of a cariogenic diet could be modified by the plaque disrupting beneficial effects of toothbrushing with fluoridated toothpaste and exposure to fluoridated water (Gibson and Williams, 1999, Harris et al., 2004, Gussy et al., 2006, Gussy et al., 2016). Children with a history of snacking once or more per day or who consumed sweets a minimum of once a day have higher rates of dental caries (Abreu et al., 2015, Bonotto et al., 2017). Bonotto et al (2017) recently found that snack limits established by parents were associated with a lower prevalence of untreated dental caries in preschool children. Obviously, the parent or primary caregiver largely controls the amount and frequency of food the preschool child may consume. Given that food and drink items consumed between the traditional main meal patterns typically contributes more than 25% of the daily energy intake of preschoolers it is surprising that the determinants of snacking at this age are not well established (Fisher et al., 2015, Jacquier et al., 2017).

Prior to water flouridation, the landmark study in Vipeholm, Sweden, in the 1940's (Gustafsson et al., 1954), concluded that frequency and timing of consumption of CF were more important than the total amount. More recent studies have shown that this relationship still remains but is not as strong (Burt et al., 1988, Sheiham, 2007, Moynihan and Petersen, 2004). Both frequency of consumption and total amount of sugars/CF are strongly related (Moynihan and Petersen, 2004) but there is still debate about which plays a more dominant role in relation to dental caries development (Moynihan, 2002, Diaz-Garrido et al., 2016). Some authors have have suggested that the evidence for both is similar (Moynihan, 2002), others that the total amount is more strongly associated with caries increment (Bernabe et al., 2015, Burt et al., 1988, Zero, 2004). A systematic review (Moynihan and Kelly, 2014) which informed the WHO guidelines on sugar intake (World Health Organization, 2015) "identified largely consistent evidence supporting a relationship between the amount of sugar intake and the development of dental caries across age groups". A recent study (Diaz-Garrido et al., 2016) examined, in vitro, the effect of increased

sucrose exposure on the cariogenicity of a biofilm-caries model and concluded that even one daily exposure to sugar can initiate a carious lesion. Although confirmation would require clinical trials to account for other protective factors such as saliva, the results suggested that higher exposure to sugar increases the acidogenicity and virulence of the Streptococcus mutans biofilm in a frequency-dependant manner. Establishing what the threshold is for the number of "exposures" and/or frequency of intake of CF per day to induce ECC is difficult but has obvious implications for health education and advice on caries risk. Other factors affecting cariogenicity include the physicochemical properties of the food item (e.g. texture, stickiness), sequence and timing of consumption, duration in the mouth and physical and chemical interactions with other food components (Moynihan, 2002, Marshall et al., 2005, Lingström, 2009, Bradshaw and Lynch, 2013).

In nutritional epidemiology there has been considerable focus, and controversy, on the role of "unhealthy" and "high- sugar" food and beverages in relation to obesity (Hall et al., 2015, Erickson et al., 2017, Erickson and Slavin, 2015b, Farpour-Lambert et al., 2015, Hu, 2013, Lustig et al., 2012) but comparatively little emphasis on assessing if the pattern of consumption of CF can be further defined to optimise dietary advice and reduce the risk of ECC (Johansson et al., 2010, Marshall et al., 2005, Amezdroz et al., 2015, Chaffee et al., 2016). People tend to consume combinations of food items as snacks or main meals rather than individual nutrients (Leech et al., 2015) and recent research has focussed on meal composition and eating patterns to explore diet-disease relationships (Faber et al., 2013, Golley et al., 2017). There are many obstacles to investigating consumption patterns of preschool children including instrument selection, proxy reporting and seasonal variations in consumption (Magarey et al., 2011). Even accepting the limitations of measuring food intake in young children there is no universal, standardised method for reporting the data and comparisons between studies can be very difficult (Faber et al., 2013). Instrument selection and study design can impact on the quality of data collected and interpretation of results is often dependant on how the data is reported. For example, Faber commented that a study on young children in South Africa sugar was ranked 17th based on the total amount consumed yet 3rd based on most frequently consumed (Faber et al., 2013).

At the individual level there is even less standardisation of dietary assessment and recent surveys of dentists' use of food diaries suggested that further research is required to develop a more reliable, acceptable dietary assessment

tool for use in the dental setting (Arheiam et al., 2016a, Arheiam et al., 2016b). Dentists are required to provide dietary advice to patients to promote good oral health and reduce the frequency and amount of potentially cariogenic foods and drinks (Moynihan, 2002). However, this practice is not standardised and there is no universal dietary assessment system to monitor patient diets (Dye et al., 2004, Moynihan and Kelly, 2014).

Professional recommendations are that the first dental visit should occur at 12 months of age, primarily to offer preventive advice to parents regarding inappropriate dietary behaviours (American Academy of Pediatric Dentistry, 2016)). Furthermore, it is an opportune time to examine for early signs of enamel demineralisation ("white-spot lesion") while remineralisation and arrest of caries is still possible (Tinanoff and Reisine, 2009, Selwitz et al., 2007, Gussy et al., 2006) by utilising dietary counselling strategies (Marshall, 2013, Marshall et al., 2005, Moynihan and Petersen, 2004, Moynihan and Kelly, 2014). However, dietary recommendations need to be practical and realistic to translate into behavioural change in the context of parental education, parental diet, and socioeconomic background (Leech et al., 2015, Faber et al., 2013). Apart from identification of foods high in free or total sugars dietary counselling advice is often limited to the reduction of between-meal snacking of sugary foods and drinks (Arheiam et al., 2016b, Moynihan, 2002). Obviously, any such "reduction" is potentially beneficial in reducing caries risk but the premise of any practical advice must be an understanding of the current nutritional status of the child. Further research is required to develop a more reliable, acceptable dietary assessment tool for use in the dental setting.

The dental health risks of different foods depend on multiple factors and previous attempts to categorise them according to their relative cariogenicity have proved difficult (Curzon and Hefferren, 2001, Lingström, 2009, Zero, 2004, Johnson et al., 2016). Food and drinks that have been categorised as potentially cariogenic (Moynihan, 2002, Moynihan et al., 2009) are listed in Table 1.3. Selecting a healthy diet that is concomitant with oral health should be an important component of both population policy guidelines and individual patient counselling (Navia, 1994). The most recent food pyramid guidelines in Ireland suggest "small servings of chocolate, biscuits, cakes, sweets, crisps, ice cream and sugary drinks - not every day, maximum once or twice a week" (http://www.safefood.eu/Healthy-Eating/What-is-a-balanced-diet/The-Food-Pyramid.aspx). An 'unhealthy diet' is associated with a number of poor health outcomes and most of the items at the top of food pyramid guidelines are also

cariogenic (Food Safety Authority of Ireland, 2011b, Food Safety Authority of Ireland, 2011a).

Table 1.3 Selected foods and drink that are potentially cariogenic.

Biscuits
Cakes and pastries
Sugars, syrups, caloric sweeteners
Ice creams
Fruit juices
Sugared, Ready To Eat Breakfast Cereals (RTEBC)
Carbonated beverages (non-diet)
Squashes, cordials and fruit juice drinks
Chocolate confectionary
Non-chocolate confectionary
Tinned fruit in syrup
Rice puddings and custards
Desserts

Source: Amended from Moynihan (2002).

However, some foods that contribute significantly to nutrient intake such as RTEBC can also contain high levels of added sugar (Priebe and McMonagle, 2016). Chankanka (2015) found that between 3 and 5 years of age different foods and beverage categories were associated with increased risk of cavitated caries compared to non-cavitated caries. Furthermore, greater consumption frequency of RTEBC at meals was associated with a greater risk of non-cavitated caries while greater consumption frequency of sugary drinks and added sugar at snacks were associated with a higher risk of cavitated caries. Many studies have found stronger associations between consumption of non-diet carbonated beverages or sugar sweetened beverages and ECC than that between fruit juice, with a similar sugar content, and ECC (Marshall et al., 2005,

Marshall et al., 2003). Some researchers (Gussy et al., 2016, Sohn et al., 2006) have similarly speculated that as the non-diet carbonated beverages are highly acidic with increased buffering capacity which may result in a lower pH at the tooth surface, they may be more cariogenic. However, these properties are more likely to contribute to contribute to dental erosion which is strongly dependant on the type of acid and extent of buffering capacity and titratable acidity (Barbour, 2009, Shellis et al., 2013).

World Health Organisation (WHO) guidelines recommend that for adults and children the intake of free sugars should not exceed 10% of total energy and a conditional recommendation of a further reduction to below 5% of total energy (World Health Organization, 2015). However, some have argued that restrictive recommendations and mandatory food labelling may not be very effective methods for reducing free sugar intake (Erickson and Slavin, 2015b, Erickson et al., 2017). There have also been conflicting views about the efficacy of providing information on reducing sugar intake as a preventive measure (Lingström, 2009, Sheiham and James, 2015). It is important to understand that the recommendations by the WHO are aimed at the level of the population rather than that of the individual (Divaris, 2016, Meyer and Lee, 2015).

1.5.2. Assessing Dietary Intake

"It is easy to ask what people eat but finding an answer can be a daunting task. For one thing, people do not eat the same food every day or week, or throughout the year or their lives. Second, every person has different needs and meets them in a different way from other people" (Becker and Helsing, 1991).

The main aim of nutritional epidemiology is to describe the variation and distribution in peoples' nutritional behaviour and relate that behaviour to causes or prevention of a health outcome (Mackerras and Margetts, 2005). A key aspect is to assess how useful and valid the estimations of dietary intake are and how that may impact on investigating health outcomes and inform policy for dietary advice at an individual or population level (Rockhill, 2001, Faber et al., 2013). Dietary assessment includes food availability at a national level, food purchase at household level and food consumption at the individual level (Thompson and Subar, 2013). This thesis is only concerned with dietary intake assessment at the individual level. While there is no generally accepted instrument for capturing intake of CF the FFQ has been considered the instrument of choice for studies exploring the relationship between diet and

dental caries due to its convenience, ease of use by subject/parent and, if it questions intake over a sufficiently long period, ability to estimate habitual food intake (Llena and Forner, 2008). The various methods for collecting dietary intake data (Table 1.4) have been compared by numerous authors (Magarey et al., 2011, Biro et al., 2002, Moynihan et al., 2009, Satija et al., 2015, Shim et al., 2014, Thompson and Subar, 2013). Methods commonly available for dietary estimation in public health epidemiology include the Food Frequency Questionnaire (FFQ), 24-h recalls (24HR), multiple-day food diaries or records, diet histories and short food questionnaires (SFQ) (Golley et al., 2017, Thompson and Subar, 2013, Shim et al., 2014). More recently, biomarkers have been developed for validation of nutrient intake that highlight other person-specific errors that can occur with all self-report methods (Jenab et al., 2009).

Table 1.4 Comparison of dietary assessment methods.

Dietary method	Description	Application	Limitations
Weighed food diary	Subject weighs and records all food consumed for defined period of time.	Accurate assessment of food and nutrient intake and may apply to all types dietary data, e.g. assessment of individual intake.	Weighing food may affect food selection and usual intake, requires literacy, eating outside home problematic.
Estimated food diary	Subject records all food consumed over set number of days in diary using household measures to estimate portion size.	Suitable for looking at changes in diet over time and assessing individual nutrient intake.	Accuracy reduced due to portion size estimation, requires literacy of subject.

Precise weighing method	All ingredients, foods served and left-over food weighed and an aliquot chemically analysed for nutrient composition.	Accurate information on nutrient intake and accounts for systematic error using food tables.	High subject participation burden, prone to reduced cooperation and record error as number days increase.
Dietary history	Detailed questionnaires administered by trained interviewer, dietician on present/past intake.	Measuring habitual intake of individuals over a long period. Compared with dietary recommendations.	Memory reliant, requires skilled dietician, recall bias
24-hour recall (24HR)	Subject recalls all food consumed in previous 24h in interview.	Suitable for average population intake.	Memory reliant, no account of daily variation, not suitable for reliable data on individual dietary intake.
Repeat 24-hour recall	24-HR repeated several occasions.	Suitable for average intake estimation and range of intake populations.	Memory reliant, not suitable for absolute nutrient intakes or comparing levels of intake to dietary recommendations.
Food frequency questionnaire (FFQ)	Self-administered questionnaire in which subject	Suitable for classifying subjects into bands of intake and for relative ranking of	Not suitable for comparing levels of intake to dietary recommendations, little detail on other

	indicates their usual frequency of consumption of each food from a list of foods for a specific period.	individuals within study. Easy to apply to large surveys and low cost.	aspects, e.g. methods of cooking, portion size.
Short frequency (or food) questionnaire (SFQ)	Simplified or targeted SFQs, questionnaires that focus on specific eating behaviour, usually records small number of food groups.	Can be useful in situations that do not require either assessment of the total diet or quantitative accuracy in dietary estimates. Low cost.	Does not capture information about entire diet, measures not quantitatively meaningful, can't estimate dietary intake population.

Source: Amended from Moynihan et al. (2009).

Table 1.4 (continued) Comparison of dietary assessment methods.

The FFQ is the most widely used dietary method for large-scale studies and this is sometimes further modified in terms of time-frame, food items and estimation of quantity to a simple dietary screening tool or SFQ (Golley et al., 2017). However, the usefulness of the resultant data that can be subsequently obtained, especially from a SFQ, may be limited, particularly if key foods are omitted and minimal consumption frequencies recorded. For example, even relatively simple descriptive analysis of "unhealthy" food intake data can be compromised and bias our understanding of habitual food consumption patterns (Anderson 2016, Kirkpatrick, 2014). Furthermore, measuring typical food intake has a number of inherent issues such as self-selection and social desirability bias and selective underreporting of specific foods (Lissner, 2006, Magarey et al., 2011, Thompson and Subar, 2013). Combining different assessment instruments can improve intake estimates (Carroll et al., 2012). Combining data from a number of 24HR with a FFQ has been shown to produce

better estimates compared to a single FFQ or a few 24HR alone when exploring associations with health outcomes by modelling usual intakes (Thompson and Subar, 2013). Other approaches have developed mixed effects models for reducing the measurement error when estimating typical nutrient intake with repeated 24HR (Tooze et al., 2010). Methodological issues in research in both oral health and dietary assessment are considerable (O'Mullane et al., 2012, Thompson and Subar, 2013, Foster and Adamson, 2014, Archer and Blair, 2015). Comparisons between studies to identify the most important dietary factors in oral health is difficult due to a lack of standardisation, use of non-validated measures, brief dietary assessment instruments, reliance on parental recall of dietary habits and inconsistent reporting methods (Thompson and Subar, 2013, Harris et al., 2004, Magarey et al., 2011, Taren et al., 2006).

1.5.3. Data mapping

Linking different datasets and augmenting information in one dataset with another is another way of optimising data quality and yield from subsequent analyses (Slack-Smith, 2012, Schenker and Raghunathan, 2007). The quality of food intake estimates may be augmented by using more accurate information in another database and carrying out a data-mapping procedure (Kettler et al., 2015). This type of data mapping allows for detailed dietary data from a matched cohort to be mapped to simple data from a large cohort with the aim of improving the quality of dietary data in large cohorts. Tools such as natural language processing (NLP) can be used to sort, group and filter categories to facilitate easy mapping. Combining information from multiple surveys has been highlighted by others as an important area of research which can elicit useful information gain and augment estimates of parameters lacking in other surveys (Schenker and Raghunathan, 2007).

1.6. Sugars consumption and dental health

1.6.1. Sugars and health

In the early Islamic period sugar had a special status as a medicine. As Mintz (1985), stated in his classic anthropological study ‘so useful was sugar in the medical practice of Europe from the thirteenth through the eighteenth centuries that the expression “like an apothecary without sugar” came to mean a state of utter desperation or helplessness’. Some medical doctors proposed that sugar

was a “veritable cure-all, its only defect being that it could make ladies too fat” and even suggested that it made for a valuable dentrifice (Mintz, 1985). Nowadays, the main functionality of sugars is their sweet taste, but sugars also contribute to colour, texture, flavour, structure and shelf-life (Clemens et al., 2016).

Dietary sugars have become a focus of much debate in public health in recent years with concerns regarding increased obesity prevalence and negative impact on oral health the focus of recently updated guidelines by the WHO and the Scientific Advisory Committee on Nutrition (SACN) (World Health Organization, 2015, SACN, 2015, Pyne and Macdonald, 2016). It is generally accepted that a reduction in sugar consumption levels may improve diet quality and be beneficial to both oral and general health (Watt and Rouxel, 2012, Gibson et al., 2016, Sheiham and Watt, 2000, Meyer and Lee, 2015). While much interest prevails among regulatory agencies and government bodies to reduce the sugar content in food and diet to help curb obesity prevalence (Edwards et al., 2016, Public Health England, 2015b) this may have a greater impact on dental caries rates given the stronger evidence base for the role of sugar (Moynihan, 2016, Moynihan and Petersen, 2004, Marshall et al., 2005, Sheiham and James, 2015). Sugar does not appear to uniquely contribute to obesity above its caloric intake contribution (Kahn and Sievenpiper, 2014, Brand-Miller and Barclay, 2017, Stanhope, 2015) but some commentators have argued that sugar and sugar sweetened drinks in particular, are a key driver of obesity related diseases over the last few decades (Bray and Popkin, 2014, Cantoral et al., 2016, Hu, 2013). Others have argued that from a common risk factor approach there is sufficient epidemiological evidence to support any measures that reduce overall sugar intake (Hopcraft and Beaumont, 2016, Sheiham and Watt, 2000, Watt and Sheiham, 2012). A reduction of sugar intake across the population has long been a goal in public dental health (Sheiham and James, 2014, Nicolau et al., 2003) and now appears to be the focus of general public health advocates (Anderson, 2014, Hu, 2013, Lustig et al., 2012).

1.6.2. Terminology and classification of dietary sugars

The term “sugars” refers to all sugars. By definition, this means all monosaccharides and disaccharides in a food, whether naturally occurring or added during processing or cooking (Stephen et al., 2012) whereas “sugar”

refers to sucrose or “table sugar” (Marshall, 2015). Total sugars (TS) is the sum of natural and added sugars (AS) in a food. Intrinsic sugars can be defined as the sugars incorporated in the structure of intact fruit and vegetables (Marshall, 2015) (Figure 1.4). Milk sugars are the natural sugars present in milk. AS include those sugars added during the production or processing of food and not naturally found in the food product (Erickson and Slavin, 2015b) and is the term defined by the Food and Drug Administration (FDA) in the USA. Free sugars (FS), which is the preferred term used by the WHO, includes sugars naturally present in honey, syrups, fruit juices and fruit juice concentrates as well as AS. TS can also be considered as the sum of FS, intrinsic sugars and milk sugars. AS and FS are not chemically distinguishable from those sugars naturally occurring in food and drink.

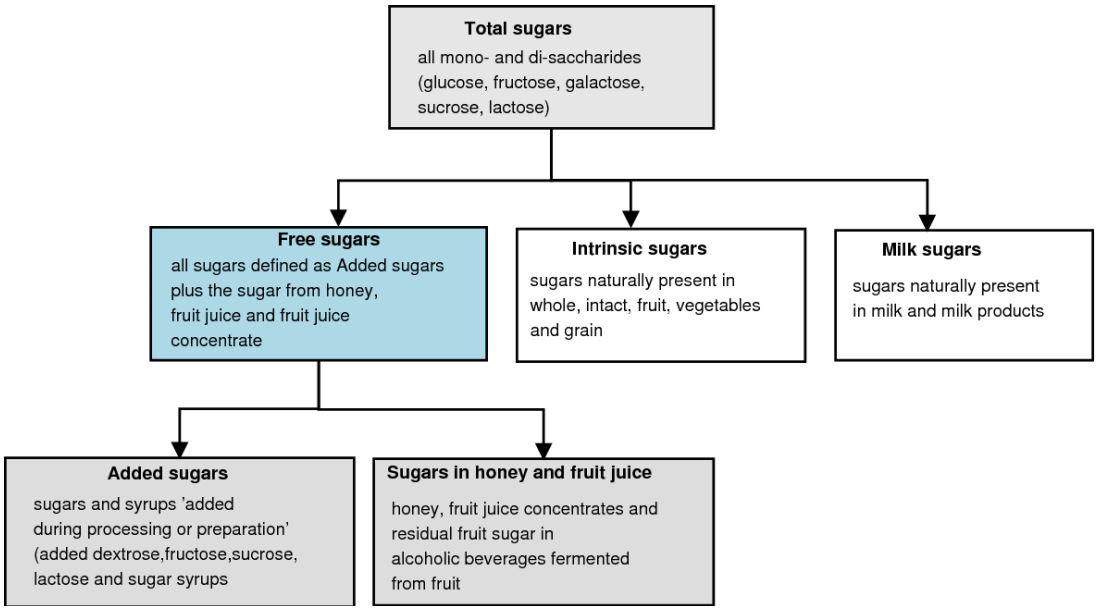


Figure 1.4 Classification of dietary sugars. Amended from Moynihan et al. (2018).

1.6.3. Dietary recommendations

Dietary FS are the most important risk factor in the development of dental caries and can contribute to excess energy intake. Consequently, the WHO have

issued recommendations to apply over the lifetime; firstly, that adults and children reduce their daily intake of FS to <10% of total energy intake (TEI) (strong recommendation) and secondly, that a further reduction to <5% of TEI could provide additional health benefits (conditional recommendation) (World Health Organization, 2015). Evidence supporting these updated guidelines were primarily based on a systematic review of the relationship between sugar intake and dental caries (Moynihan and Kelly, 2014) and that between sugar and obesity (Te Morenga et al., 2013). It is interesting that while the WHO concluded that the overall strength of evidence of the relationship between consumption frequency and amount of FS and dental caries supported a policy for a stricter limit, it also acknowledged the weakness of the evidence linking the association between FS and obesity (Moynihan and Kelly, 2014, Moynihan, 2016). Recommendations by the SACN were that the average population intake of FS should not be greater than 5% of TEI for age from 2 years upwards (SACN, 2015). The WHO have also recommended that the diet of two-year olds should have “no sugars” from the perspective of a common risk factor approach (World Health Organization, 2015).

1.6.4. Free sugar estimation and consumption

While it is vital to understand national consumption levels before considering health strategies or policy (Brisbois et al., 2014) very few food databases contain information regarding AS or FS levels (Erickson and Slavin, 2015b, Newens and Walton, 2016, Brisbois et al., 2014). This may be due to multiple difficulties in estimating AS or FS levels (Louie et al., 2015, Newens and Walton, 2016). There is a lack of agreed standards in terms of definitions, terminology, standardisation of estimation procedures and protocols for dealing with the food industry in terms of ingredients/recipe and sugar reduction programmes (Public Health England, 2017, Erickson and Slavin, 2015a, Erickson and Slavin, 2015b).

While developed countries have reported an overall reduction in consumption of AS/FS in the last 10-20 years (Welsh et al., 2011, Brand-Miller and Barclay, 2017, Wittekind and Walton, 2014) the majority of dietary research has focussed on adults and older child cohorts whereas detailed information about intake of food and drink with AS or FS in preschoolers has been lacking (Amezdroz et al., 2015, Mis et al., 2017). A recent review which examined sugar consumption in children, adolescents and adults, measured by dietary surveys worldwide, concluded that AS/FS intakes in children and adolescents are

higher than in other population groups and further research into dietary patterns contributing to this is needed (Newens and Walton, 2016). Intakes of TS in 3 year old children ranged from 24.5 to 29.0 as a percent of TEI. From a total of 10 surveys in children less than 4 years old only four reported AS intake and for 3 year olds this ranged from 7.5 to 10.6 % of TEI (Newens and Walton, 2016). One Irish study examined the quantitative relationship between the frequency (four times per day) of AS intakes and 10% TEI from AS in children (5-12 years), teenagers and adult (Joyce et al., 2008). The researchers suggested that, generally, the percentage energy from AS intakes increased as the frequency of eating occasions (of AS) increased but there were notable differences between age groups and food sources. The main contributors to AS intake in children and adolescents were biscuits, cakes, buns and pastries, carbonated beverages, squashes and cordials, confectionery and RTEBC. Other researchers have found similar trends with children and adolescents in particular, showing higher contributions of FS/AS to TEI than older cohorts and have strongly recommended reducing intakes of energy-dense-nutrient poor foods and drinks (Lei et al., 2016, Ruiz et al., 2017, Gibson et al., 2016, Brisbois et al., 2014, Sluik et al., 2016).

1.7. Data mining and data analysis techniques

“We are drowning in information and starving for knowledge”

(Rogers, 1985)

Data mining is a technique used in data analysis which has been described as the discovery of models (Leskovec et al., 2014) and also as “the art and science of intelligent data analysis” (Williams, 2011). Previously, the term was synonymous with “data dredging” which was a derogatory description of dubious efforts to extract (“torture”) from the data information that was not supported by it. However, the current view is that data mining involves the development “of a statistical model, that is, an underlying distribution from which the visible data is drawn” (Leskovec et al., 2014). Data mining utilises techniques derived from computer science, machine learning and statistics. Models can help in the understanding of real world concepts but can also be used to make predictions. George Box’s famous aphorism stated, “all models are wrong, but some are useful” (Box, 1976). While developing a model may

allow one to interpret the data in a useful way it is necessary to then evaluate and, potentially, deploy the model to deliver a benefit (Williams, 2011). However, the focus is removed from utilising a “regression approach” to all epidemiological investigations as was best stated in a recent essay by Rothman (2017): “Today it is typical to train epidemiologists to use regression models as their first approach to data analysis. Such training fosters the idea that regression modelling is the primary, and perhaps the only approach to use in analysing epidemiologic data. As a result, the rift between epidemiologists and their data, and more so between readers and the data, is growing”. The continuing discourse about null hypothesis significance testing and the “maligned” p-value has been highlighted by data scientists as an opportunity for better education across all disciplines (Leek and Peng, 2015a).

1.7.1. Data pipeline

What is the optimal approach to exploring a new dataset? What are the first steps required before beginning to build a model or carry out a formal statistical analysis? The process of data analysis involves many features but, graphical techniques are an essential step in checking assumptions, models and communicating findings (Wickham and Grolemund, 2016). Data analysis or data analytics can be described as the operations in a pipeline (Figure 1.5) which include importing, “tidying” (cleaning) and transforming data. This is followed by visualising, modelling and communicating data with the goal of discovering useful information, forming conclusions, and supporting decision-making (Wickham and Grolemund, 2016). Importing, tidying and transforming are collectively described as wrangling.

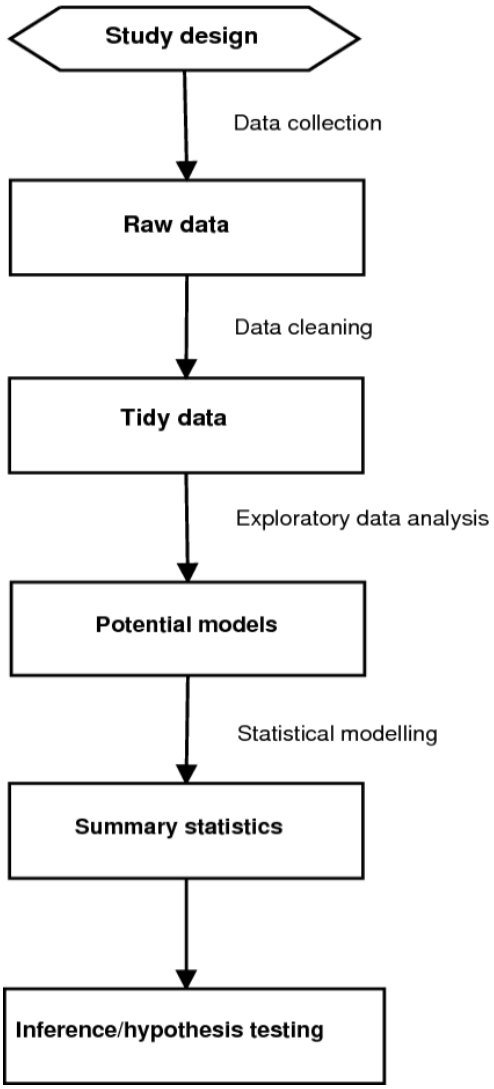


Figure 1.5 Data pipeline in the design and analysis of a successful study. Adapted from Leek and Peng (2015a).

A data pipeline can be viewed as a logical series of processing steps (inputs) that are connected together to produce an outcome (outputs) (Maimon and Rokach, 2009). The terminology in this pipeline is adapted from Leek and Peng (2015a).

While data analysis has multiple applications in diverse domains including business and science, in the area of statistical applications it commonly involves exploratory data analysis (EDA) or confirmatory data analysis (CDA). John Tukey formalised and developed the use of graphs to display and understand data to “encourage” the data to speak for themselves as part of, or before, a formal statistical analysis (Maindonald and Braun, 2010). In CDA a

statistical hypothesis test is used, and a statistical inference approach adopted. However, depending on the aims of the study, most data analysis strategies need both an EDA and more formal analyses ((Maindonald and Braun, 2010). A key aspect is to use a step-by-step process or "Data Pipeline" that encompasses the core activities of data analysis such as: (1) Stating the question; (2) Exploring the data; (3) Building formal statistical models; (4) Interpreting the results and (5) Communicating the results (Wickham and Grolemund, 2016, Peng, 2012). There are numerous comprehensive reviews of data mining and EDA available (Williams, 2011, Leskovec et al., 2014, Hastie et al., 2009, Chawla, 2010, Yoo et al., 2014).

An important part of the data science approach is to carefully record any programming code or syntax used in any part of the pipeline. For example, R is a free open-source programming language for statistical computing and visualisation. R studio (http://www.rstudio.com/download) is a powerful integrated development environment for R (Williams, 2011). R markdown combines the code and script editor and is an excellent method for retaining both embedded code chunks and plain text formats as an accurate, reproducible record of the data analysis (Lander, 2014).

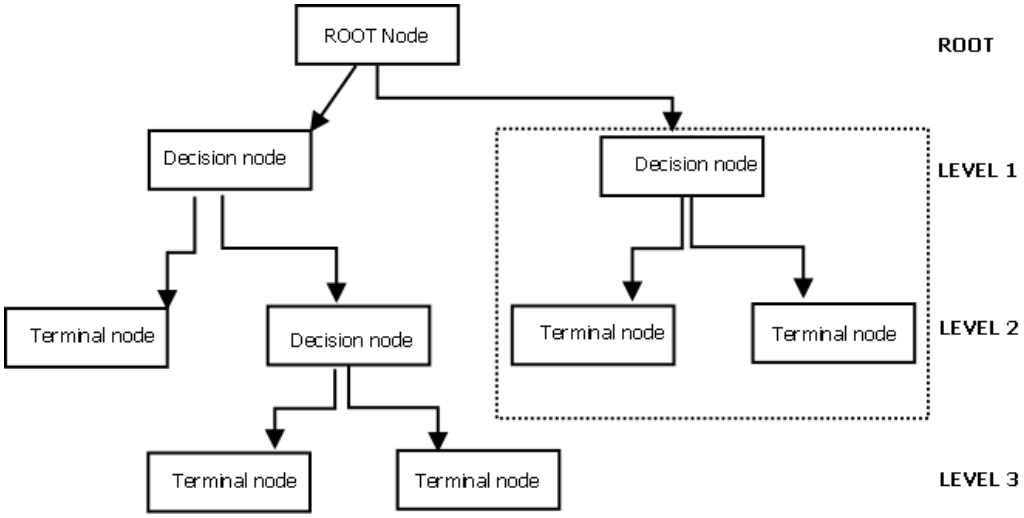
1.7.2. Learning from the data: decision tree methods

'Big data' is a term which refers to a wide variety of applications and includes many different computational and statistical approaches (Binder and Blettner, 2015). It also embraces the concept of increasing amounts of data that cannot be easily analysed using traditional methods. It is important to be aware of, and understand, the function of data mining algorithms before applying them to health outcome research (Yoo et al., 2012). Learning from the data can be subdivided into two main approaches, supervised and unsupervised. Supervised learning aims to predict an outcome or target value using a range of predictor or independent variables while the goal of unsupervised learning is to describe patterns and associations in a set of input values. Association rules and cluster analysis are examples of unsupervised learning while decision trees and linear regression techniques are examples of supervised learning (Hastie et al., 2009).

Decision tree methods are particularly popular in the data mining community for classifying multiple covariates or developing prediction algorithms for a

target variable (Song and Ying, 2015). Decision tree methods can be used for regression and classification problems. Regression trees are used for a continuous outcome variable while classification trees are used when the outcome variable is categorical or binary (Hastie et al., 2009). Classification and regression trees have been used in clinical settings for risk assessment or diagnostic prediction but less so in public health research (Yoo et al., 2012, Song and Ying, 2015, Kuhn et al., 2014, Lemon et al., 2003). The use of decision tree techniques for investigating chronic disease outcomes and dietary intake patterns has been recently described as a useful "data-driven, outcome-dependant" emerging method (Krebs-Smith et al., 2015).

The aim of classification tree analysis (CTA) is to create a model that predicts a target outcome (dependent variable) based on the strength of interactions between categorical or continuous input variables (independent variables) (Loh, 2014). A decision or classification tree is a flow-chart-like structure (Figure 1.6) with internal nodes that represent a test on an attribute and leaf nodes represent class labels or class distribution (Rokach and Maimon, 2009). The branch represents the outcome of the test and each leaf node represents a classification or decision. In turn, each leaf becomes a candidate to be the node for a new split. The root or parent node has no incoming edges while all other nodes have one incoming edge. The tree can be considered as a series of "nested-if" logical statements, in programming terms, which are mutually exclusive and exhaustive (Ho Yu, 2010). The most important input variables are identified and displayed at the top level of the tree and splitting of data into subgroups continues until the tree is stopped by user defined stopping criteria (Rokach and Maimon, 2009). Ultimately, CTA allows for multilevel interactions between predictor variables by using recursive partitioning of the data space (Camp and Slattery, 2002). This is regarded as a distinct advantage over more conventional methods such as logistic regression where predictor variables act independently, and interactions are usually omitted in the final model to reduce complexity. Another advantage of CTA over more traditional regression techniques is that the output produces a visualisation of all the significant interactions with the target variable at multiple levels (Kuhn et al., 2014, Song and Ying, 2015). Furthermore, as many of the new EDA techniques, such as CTA, are non-parametric, they are not subject to the assumption of absence of multicollinearity for independent variables that is required for regression analysis (Ho Yu, 2010).



- Root Node**:** represents entire population or sample and is further divided into two or more homogeneous sets.
- Splitting: It is a process of dividing a node into two or more branch nodes.
- Decision Node: All other nodes are called leaves and are represented by Terminal or Decision nodes. When a sub-node splits into further sub-nodes, then it is called decision node.
- Terminal Node**:** Nodes that do not split are called Terminal nodes.

Figure 1.6 Graphical illustration of the components and structure of a decision tree. Amended from Rokach and Maimon (2009).

Inducers of decision or classification trees are algorithms that build a tree from a dataset (Rokach and Maimon, 2009). The algorithms construct the decision tree in a top–down, recursive manner and are greedy by nature (Rokach and Maimon, 2009). Algorithms available for building decision tree models include Chi-squared Automatic Interaction Detection (CHAID) (Kass, 1980), C4.5 (Quinlan, 1993) Classification and Regression Trees (CART) (Breiman et al., 1984) and Quick, Unbiased, Efficient, Statistical Trees (QUEST) (Loh and Shih, 1997) are comprehensively reviewed by Loh (2014). CHAID, one of the most popular techniques, uses Bonferroni-adjusted chi-square statistics for split selection. Originally, the primary purpose of CHAID was to detect variable interactions, although now it is mainly used as an alternative to multiple regression in exploratory analysis especially when the data does not conform to strict parametric requirements (Miller et al., 2014).

Advantages of decision trees include that it is a non- parametric method with no assumptions about the distribution, provides graphical representations which are easy to understand, can handle missing data and can use both categorical and continuous data (Rokach and Maimon, 2009). The main disadvantage of decision trees is a tendency for overfitting which can limit the generalisability of the model (Rokach and Maimon, 2009). Tree models can be cross validated, to minimise overfitting of the model, by dividing the sample into subsamples and trees generated with the data from each subsample excluded in turn. Simply put, cross-validation divides the dataset into training and testing sets (Ho Yu, 2010, Rokach and Maimon, 2009). The first tree is generated on all cases except those in the first subsample, the second tree based on all except those in the second subsample etc. Misclassification risk can be calculated for each tree by applying the tree to the excluded subsample and the final risk estimate based on the average for all trees.

A confusion matrix (also called an error matrix) is a common method for evaluating the model performance. The matrix compares decisions made by the model with the actual decisions showing the number of cases classified correctly and incorrectly for each category of the dependant or target variable. Measures can be determined for evaluating the accuracy of the predictions or classification of the model such as sensitivity (true positive rate), specificity (true negative rate) and overall accuracy (total number of correct predictions divided by the total number of the dataset) (Rokach and Maimon, 2009).

1.7.3. Association analysis

Association analysis is an area of data mining that has grown rapidly in recent years and is a useful methodology for discovering interesting relationships hidden in a dataset. An association analysis algorithm generates frequent item sets (Williams, 2011). These relationships or correlations are expressed as a collection of association rules due to their frequent co-occurrence. Association rules have been traditionally applied in market basket analysis but more recently have also been applied to other domains such as bioinformatics, medical diagnosis and scientific data analysis (Höppner, 2005). For example, association rule mining has been used to generate quantitative rules that predict mental illness based on psychological factors (Cheng et al., 2014).

1.7.4. Visualisation of data

Given the rapid changes in data science we now have many exciting new ways to help convey the information embedded in data. Data visualisation refers to the coding techniques used in the visual representation of information obtained from data (Yoe, 2011). The importance of statistical graphics in data science was classically portrayed by Tufte in "The Visual Display of Quantitative Information" (2001). Tufte used the famous Anscombe quartet to demonstrate that graphics can be more revealing than conventional statistical parameters where all four datasets described by the same linear model have very different distributions (Figure 1.7). He stated "Furthermore, of all methods for analysing and communicating statistical information, well-designed data graphics are usually simplest and at the same time the most powerful".

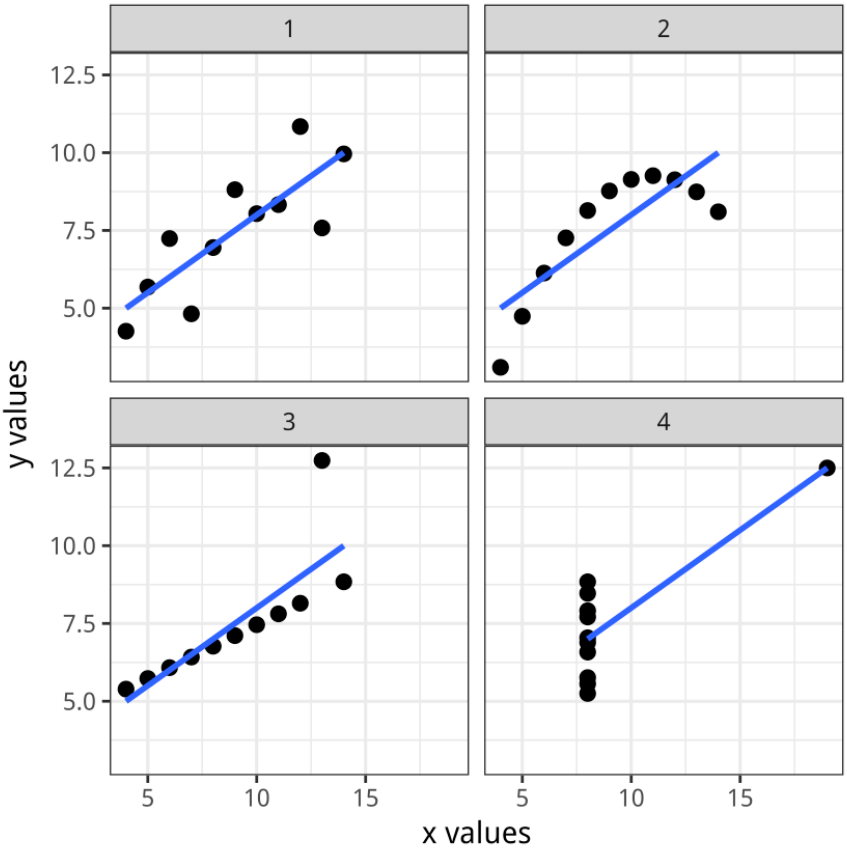


Figure 1.7 Graphs in statistical analysis. Amended from Tufte (2001).

Data visualisation has become a key component of EDA and there are many techniques that are being used more widely in the era of “Big Data”. There are many known visualisations that are used to show data patterns such as histograms, stem-and-leaf-plots, boxplots, density traces, bean plots and alluvial diagrams (Williams, 2011). Bean plots are an “alternative to the boxplot for visual comparison of univariate data between groups” (Kampstra, 2008) (Figure 1.8). The bean plot uses a combination of non-parametric kernel density estimates of the probability density function with a scatter plot of all data points. A kernel density estimates the probability density function of a random variable in a non-parametric fashion (Williams, 2011). The individual observations are depicted as short horizontal lines in a one-dimensional scatter plot and the estimated density shape of the distributions is displayed as a polygon. The name derives from green beans as the density shape represents the pod and the scatter plot shows the seeds inside the pod. The heavy solid horizontal line represents the average for each subgroup and the dashed horizontal line is the overall average value for the dataset.

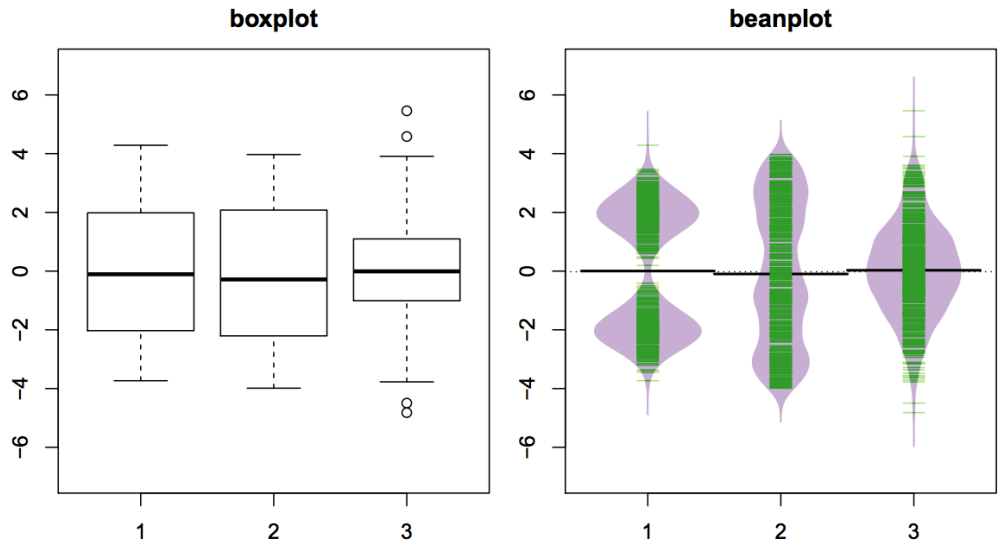


Figure 1.8 Comparison of boxplots and beanplots for a biomodal, a uniform and a normal distribution. Green lines in the beanplots show individual observations, while the purple area shows the distribution. Amended from Kampstra (2008).

Parallel co-ordinates are a multi-dimensional visualization tool first developed in 1885. Inselberg (1985) promoted the method and explained that “By means of Parallel Coordinates planar "graphs" of multivariate relations are obtained.

Certain properties of the relationship correspond to the geometrical properties of its graph. On the plane a point ←→ line duality with several interesting properties is induced". They are commonly used to identify how variables or groups may be related to each other and can represent common scales on parallel axes. This form of measurement, i.e., position along a common scale, is the easiest visual comparison for humans to make (Cleveland and McGill, 1985). A series of parallel axes are labelled according to the variables and a point on the Cartesian plane is represented by the variable value on its respective axis (Figure 1.9). Finally, the resulting points are joined by a broken line segment representing all the observations in a dataset (Wegman, 2003).

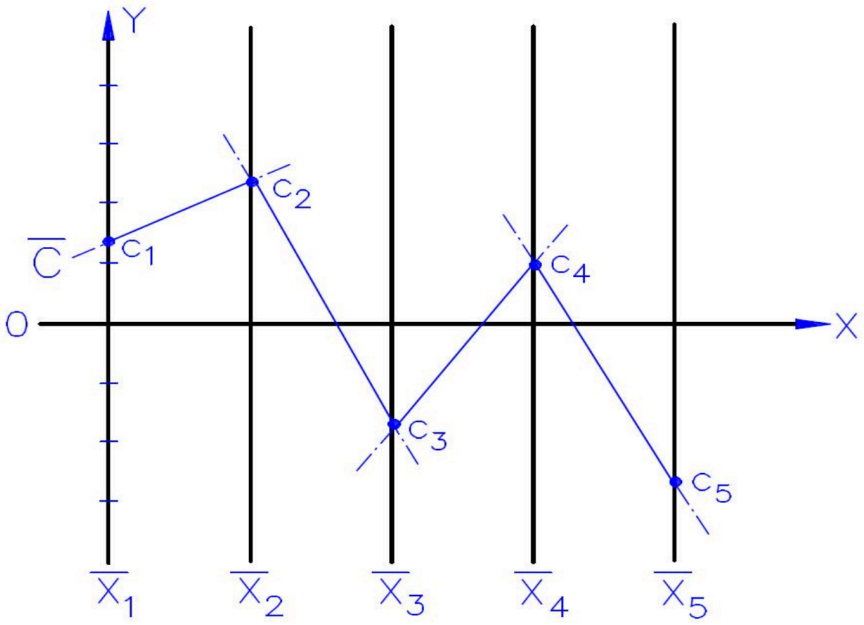


Figure 1.9 Parallel coordinates: illustration of mapping on the plane with x y-Cartesian, starting on the y-axis, N copies of the real line, labelled $x_1, x_{2...}, x_N$ are placed equidistant and perpendicular to the x-axis. The point C = (c1, c2, c3, c4, c5) is represented by the polygonal line shown. Amended from Inselberg (1985).

An alluvial diagram is a form of parallel coordinates plot used for categorical variables (Figure 1.10). They are useful for showing how multiple groups, represented as 'nodes' on a vertical axis, are related to one another over time. As with parallel coordinates, variables are positioned on vertical axes that are parallel. Values are represented by blocks on each axis and the width of the

connecting ‘streams’ illustrates the proportion in each category (Rosvall and Bergstrom, 2010).

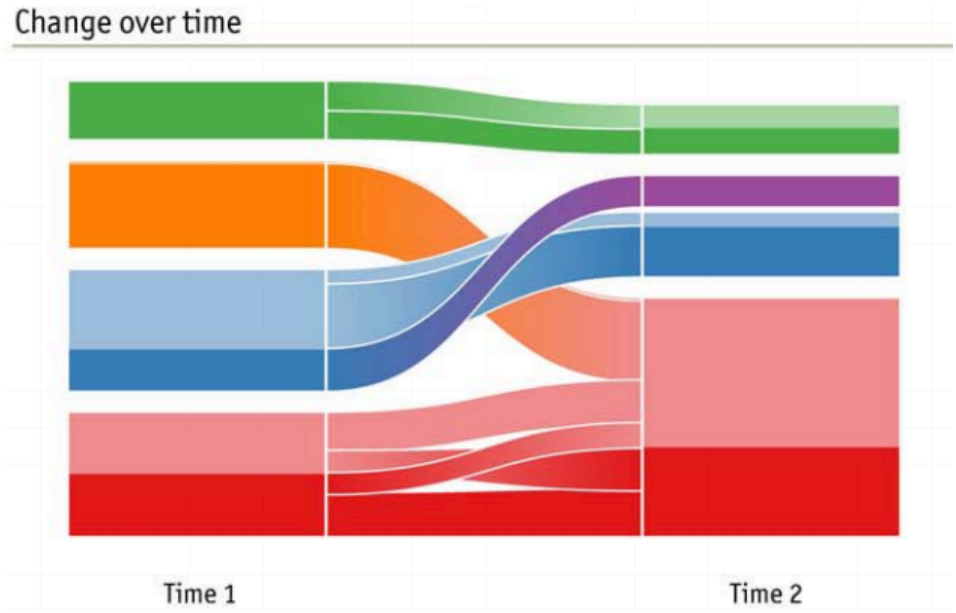


Figure 1.10 Alluival diagram mapping change in networks. The height of each block represents the volume of flow though each cluster, with significant subsets in darker colour. Amended from Rosvall and Bergstrom (2010).

1.8. Conclusions

The importance of adopting an interdisciplinary approach to oral health research in preschool children has become more evident in recent years. High quality oral health related data for the preschool age group is lacking and there are many additional complicating factors that contribute to interdisciplinary research difficulties such as proxy parent/PCG reported information, lack of standardisation of dietary intake measurement and reporting, practical difficulties carrying out clinical examinations, lack of national reference growth curves for determining weight status and ethical issues with data sharing and linkage of databases. Innovative new approaches which utilise the powerful tools of data science can assist in progressing oral health research and effect strategies and policy development (Binder and Blettner, 2015, Divaris, 2016, Krebs-Smith et al., 2015).

The growth of 'Big Data' science has led to wider availability of these techniques but a greater understanding of both data mining and its limitations is required to successfully apply these powerful analytical tools. A strong conceptual basis is still required in oral health research as the accuracy of predictive modelling continues to improve (Binder and Blettner, 2015, Ho Yu, 2010). However, there are opportunities to successfully use a data science approach for both exploratory (hypothesis generation) data analysis and modelling (Ho Yu, 2010, Wickham and Grolemund, 2016), particularly for focusing on high risk groups such as the overweight/obese who may share common risk factors with dental caries (Divaris, 2016, Chi et al., 2017). Using a data science approach and techniques such as decision trees, it is possible to combine the "two cultures" from both the statistics and data mining communities (Breiman, 2001).

Few studies have examined patterns of consumption of CF in preschool children. It is remarkable, given the importance of CF, and free sugar specifically, in contributing to ECC that greater emphasis has not been placed on understanding these meal/snack patterns of consumption and interactions between food components. A lack of standardised methods for measuring and comparing food intake at this age and a poor level of interdisciplinary cooperation has not assisted progress in this vital area of understanding early childhood dental problems. Given the number of risk factors associated with both obesity and dental caries it is logical, from a public health perspective to pursue this common risk factor approach while acknowledging that the empirical evidence does not yet support it. This research hopes to encourage an interdisciplinary data science approach and utilisation of predictive modelling tools to explore risk indicators and dental problem visits in children.

1.9. Aims and Objectives

The main aim of this thesis is to explore risk indicators associated with dental problem visits in preschool children in Ireland using a data science approach.

Objectives:

- Identify key associations between infant/child and PCG general, psychosocial and physical health parameters and dental problem visits using decision tree methods.

- Visualise the multilevel relations between weight status and dental problem visits in 3-year old children using classification tree analysis.
- Use data mapping protocols to compare and augment dietary intake estimates from the National Preschool Nutritional Survey (NPNS) database and Growing Up in Ireland (GUI) cohort survey.
- Link data from NPNS and GUI infant surveys to investigate the distribution patterns of cariogenic food and drink consumption in preschool children combined with meal association analysis.
- Determine and describe the consumption pattern of dietary free sugar in 3-year old Irish preschool children in relation to recent dietary goals/recommendations by the WHO.

Chapter 2. General Methods

2.1. Overview

A key theme of this research study was to use data science analytical techniques to investigate risk factors/indicators related to dental problems in child cohort surveys, particularly, the Growing Up in Ireland (GUI) study. Most analytical approaches to national cohort surveys have previously focussed on reporting parameters such as population mean or median and used regression models for investigating associations between dependent and independent variables (Christensen and Langberg, 2012) This thesis focussed on using descriptive analytics and predictive models (supervised learning) rather than discovery models (unsupervised learning). The traditional focus of modelling is inference or hypothesis confirmation (Wickham and Grolemund, 2016) using approaches such as null hypothesis significance testing (NHST) (Leek and Peng, 2015a). While there is much ongoing controversy about the value of NHST and p-values, decisions taken earlier in the data pipeline have, arguably, greater impact on the end results (Leek and Peng, 2015b). In this study, a data science approach was adopted to explore the question of factors affecting child dental problem visits in preschool children. The cycles of data acquisition, setting the question, exploratory data analysis, model building, interpretation and communication were the framework for this investigation (see Figure 1.4) (Zumel et al., 2014, Peng and Matsui, 2015). Interpretation of initial models also included acquiring additional data from other sources, specifically, the National Preschool Nutrition Survey (NPNS) and continuing the same theme of data science analysis. Part of this approach included using the R programming language for statistical computing and visualisation and reporting while SPSS was used in the early stages of the study and for decision tree building.

The initial phase of this research focussed on conceptual models of oral health and associated risk indicators in the GUI infant cohort of preschool children. The second phase concentrated on the interdisciplinary aspect of dental problems and weight status with dietary intake as a common risk factor. Finally, analysis focussed on linking the NPNS and GUI datasets to explore the dietary risk factors in more detail including free sugar consumption, cariogenic food intake patterns and meal analysis. The following sections (2.2 and 2.3) provide details of the data sources and analytical methods used and their application

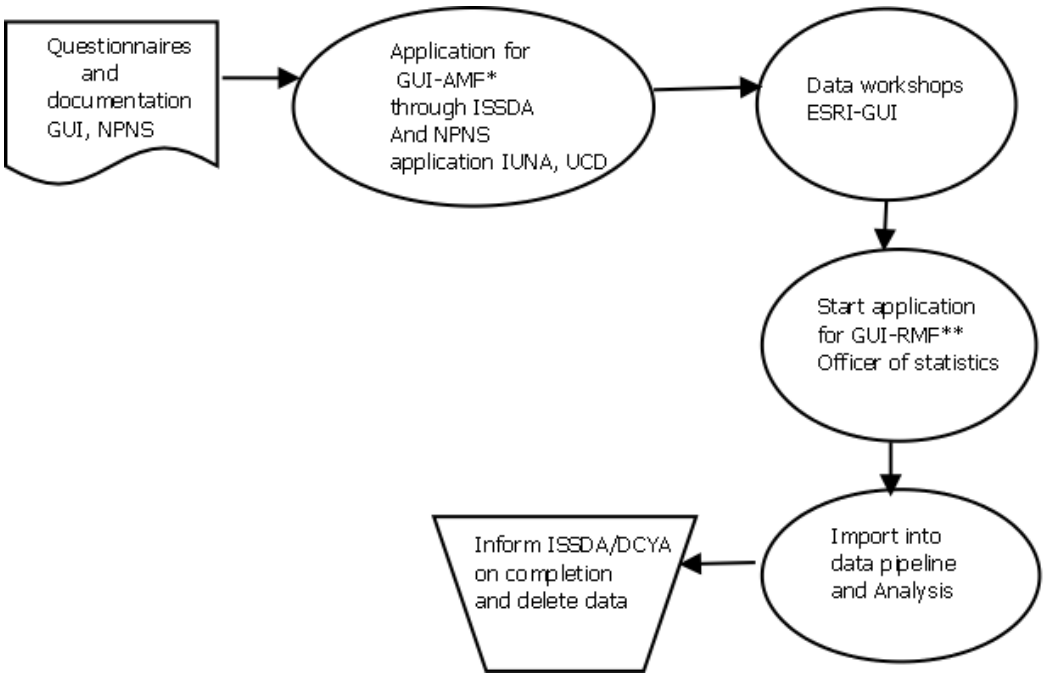
to each chapter is outlined in detail in sections 2.4-2.6. Relevant code fragments are provided in the methods section of each chapter or code used for analysis (SPSS/R) listed in “index” Table A1 and included in Appendix A.

2.2. Data sources and the study population

‘…invariably, simple models and a lot of data trump more elaborate models based on less data’ (Halevy et al. 2009)

Data sources used were derived from the GUI survey (http://www.esri.ie/growing-up-in-ireland/) and the NPNS (http://iuna.net/index.php/national-pre-school-nutrition-survey-npns/). An overview of each cohort was described in Chapter 1 and further details of the study participants, data collection and data pipeline are outlined in each section below. The main datasets utilised were derived from the infant cohort of the GUI study and the NPNS study.

Before importing the datasets in to the pipeline the following steps were taken in the initial phase of accessing the data:



* Anonymised Microdata Files (AMF)

** Researcher Microdata Files (RMF)

Figure 2.1 Data access protocol for Growing Up in Ireland (GUI) and the National Preschool Nutrition Survey (NPNS) datasets.

The initial datasets obtained from the Irish Social Science Data Archive (ISSDA, University College Dublin) were described as the Anonymised Microdata Files (AMF) and were deposited as flat rectangular data files in SPSS format. The information contained in the file was collected under the Statistics Act, 1993 and can only be used for statistical purposes. Subsequently, more detailed Researcher Microdata Files (RMF) were obtained after appointment of the researchers as Officers of Statistics by the Director General of the Central Statistics Office (CSO). The RMF files contained less collapsing and more variables but had more restricted access. There were three main categories of information available on the RMF only; identifiable information such as details of chronic physical health problem or illness, sensitive information such as variables related to survey questions on drug-taking or alcohol abuse and, finally, individual items from instrument scales. Initial analysis in this research study used the AMF but all results presented were from analyses using the RMF.

There were two types of statistical adjustment factors included in the data files: a weighting factor and a grossing factor. The weighting factor was applied to the total number of children in the GUI sample and generally used for descriptive analysis while the grossing factor grossed to the total number of children in the Irish population to provide population estimates. Prior to analysis, Wave 1 data were reweighted based on population statistics from the Central Statistics Office (CSO) and the National Child Benefit Register for 2008 to ensure that sample was representative of the total population. Data from Wave 2 were adjusted for the children who were resident in Ireland at Wave 1 but not at Wave 2, and for differential response and attrition between interviews (Williams et al., 2013). As our primary interest was to investigate the relationships between these predictor variables and dental problems at a population level, all analyses of the GUI infant cohort reported here were carried out with weighted data.

The National Child Benefits Register was used as the sampling frame for the GUI infant cohort which is a universal welfare entitlement in the Republic of Ireland. A random sample of over 11,134 households was recruited using a simple, systematic selection procedure, a random start and a constant sampling fraction (Quail et al., 2011). The sample was pre-stratified by marital status, county of residence, nationality and number of children. Families were first invited to participate when the infant was 9 months old in 2007 (Wave 1). The sampling fraction was 0.42 with an overall response rate was 64.3% at Wave 1. Subsequent follow-up interviews took place between December 2010 and July 2011 (Wave 2) when the children were 3 years old. Family response was 91% at Wave 2 while 3.8% emigrated or deceased and the remainder were either uncontactable or refused to participate. The GUI survey used a fixed panel design. Full details of the population, sample design, participant response, fieldwork/implementation, survey instruments, structure and content of the datafile and interviewer training are available from GUI at http://www.esri.ie/growing-up-in-ireland/ (Quail et al., 2011, Murray et al., 2013).

For the GUI study, interviews were carried out by trained fieldworkers administering a computer based questionnaire with the primary caregiver (PCG) in the family home after informed consent was obtained in writing (Williams et al., 2013). The PCG was defined as the person, in most cases the mother and biological parent, who delivered most care to the study child and who was best placed to provide information about him/her. The Computer-

Assisted Personal Interview (CAPI) questionnaires used in GUI mainly used closed questions. The program incorporated an extensive range of cross-variable consistency checks (Murray et al., 2013).

The NPNS had a total sample of 500 children aged 2-4 years; but only the 3-year olds were included for this analysis (n=126). NPNS used a quota sampling approach to obtain a nationally representative sample of 125 children within each of the four preschool age groups between 1-4 years of age (Walton and Flynn, 2013). The NPNS sample was recruited from an Irish parenting resource database (https://www.eumom.ie/) or from childcare facilities randomly chosen in selected locations (Irish Universities Nutrition Alliance, 2012). The NPNS datafiles are available on application to IUNA. Field work for NPNS started in October 2010 and was completed in September 2011. In the NPNS study the researcher visited the participant's home on three occasions during the 4-day food record period. Full details for NPNS are available at http://www.iuna.net/. These include details of the quality procedures that were used to help consistency and minimise error throughout the collection and manipulation of the food intake data.

Data formats in GUI were flat rectangular files in SPSS. Datafiles in NPNS were in .csv format. GUI data files were either read directly into SPSS or R or converted into .csv format before import into R as a data frame. Both studies were conducted according to guidelines laid down in the Declaration of Helsinki. Ethical approval for the GUI project was received from a Research Ethics Committee convened by the Department of Health and Children while approval for the IUNA-NPNS project was obtained from the University College Cork Clinical Research Ethics Committee of the Cork Teaching Hospitals, University College Cork.

2.3. Data pipeline and Analytical Techniques

2.3.1. Data pipeline

The data pipeline used (Figure 2.2) was a logical series of processing steps as adapted from Wickham and Grolemund, 2016 (Wickham and Grolemund, 2016).

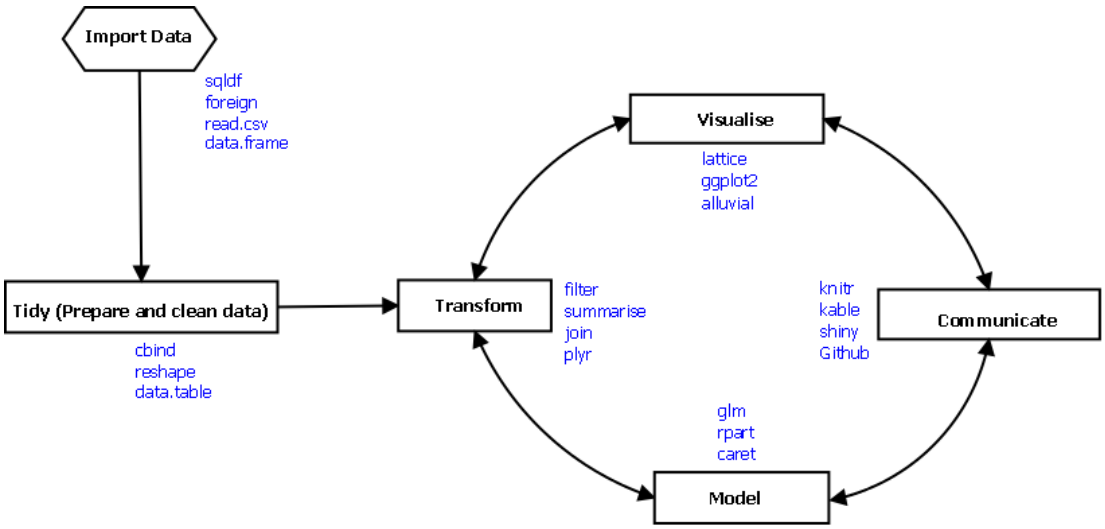


Figure 2.2 Data pipeline, indicating R packages for each step. Amended from Wickham and Grolemund (2016). R packages highlighted in blue font can be applied at each stage of the analysis pipeline.

Full details of the R language definitions and packages are available at https://cran.r-project.org/doc/manuals/r-release/R-lang.pdf. Wrangling the data is a term used to describe the functions of the packages in the pipeline that import, tidy and transform. Visualisation and modelling are the final part of the pipeline before communication or hypothesis generation.

2.3.2. Statistical programmes

R studio is an integrated development environment (IDE) for the R language. As stated on the Comprehensive R Archive Network (CRAN) R is “a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc”. All downloads of the base system, packages manuals and many other resources are available at https://cran.r-project.org/. R markdown is a variant of Markdown that has R code chunks along with Markdown text that can be rendered into a formatted html, pdf or Word file. A key part of the open data science approach includes accurate recording of the syntax used in the pipeline to provide reproducible reports and offer repeatability of the analyses (Zumel et al., 2014, Maindonald and Braun, 2010, Peng, 2012).

Initially, data cleaning, tidying and transformations for the early data in GUI was carried out using IBM SPSS statistics (v. 20.0: SPSS, Chicago, IL). Further data wrangling was carried out with R Studio (https://www.rstudio.com/). Data in SPSS file formats were either read directly or converted to .csv files before importing into R. Merging of datafiles was carried out with both SPSS and R Studio. IBM SPSS Modeler (v. 14.2: Chicago, IL) was also used for classification tree modelling. The R environment is capable of reading and processing several data formats. The functions in the readr package mainly relate to turning flat files into data frames which are the primary data structure in R. The data frame generally relates to "tabular" data, i.e. data stored in columns and rows where each column contains measurements on only one variable and each row contains one case (Wickham and Grolemund, 2016).

Raw data was checked for errors (whether in SPSS format or .csv) such as incorrect data type where numbers were stored as strings or wrong category labels. Datafiles were initially analysed using SPSS/SPSSmodeler but further analyses were carried out primarily in R studio. Many of the R packages now available for cleaning, processing and manipulating data such as tidyr, tibble and dplyr were not released when this analysis was initially completed. For example, the sqldf package was used frequently to aggregate data whereas dplyr is more commonly used now to make manipulation of data frames easier. Tibbles are enhanced data frames which are generally used currently instead of the data.frame package as they make it easier to work in the tidyverse as described by Hadley Wickham who developed the tidyr package (Wickham and Grolemund, 2016). Both SPSS and R Studio were used for variable recoding, transformations, assessing data distributions e.g., for normality, checking outliers and generating initial data visualisations and descriptive statistics. Assumptions were checked before applying any parametric or non-parametric tests such as checking residuals or inspecting correlation coefficients for multicollinearity. The analytical techniques used are outlined below and the relevant code snippets or R markdown documents detailed in each chapter and included in Appendix A.

2.3.3. Analytical techniques

Decision tree methods, specifically classification and regression tree analysis, were carried out using both SPSS Statistics and SPSS Modeler. The precise criteria and validation settings used for each analysis are listed in the relevant sections, but classification trees were typically generated using Chi-squared

Automatic Interaction Detection (CHAID) algorithm (Kass, 1980) and the Bonferroni-adjusted chi-square statistic. Binary logistic regression analyses (Agresti, 2007) were carried out using SPSS.

Model performance was assessed using a confusion matrix, also described as an error matrix (Table 2.1) to provide an estimation of selected metrics. This is a commonly used means for evaluating the quality of a model when predicting a categorical target as occurs with binary classification (Williams, 2011, James et al., 2013). Derivations from the matrix that can be used for this purpose include sensitivity or True Positive Rate (TPR), specificity or True Negative Rate (TNR) and accuracy (TP + TN/ P + N) where P= the number of real positives in the data and N= the number of real negatives in the data.

Table 2.1 Confusion matrix for model performance evaluation.

Total	Positive	Negative
Predicted positive	True positive (TP)	False positive (FP) (type 1 error)
Predicted negative	False negative (FN) (type 2 error)	True negative (TN)

TPR= TP/P = TP/ TP + FN

TNR= TN/N = TN/ TN + FP

For all statistical hypothesis tests the alfa-level was set at $p<0.05$ or $p<0.01$. Statistical tests used to compare differences between groups included:

1. The chi-square test for homogeneity (or test of two proportions) was used to determine if there was a difference in proportions between two groups ($p<0.05$ alfa level) (Woodward, 2013).
2. Equivalence tests were carried out to determine if population means were equivalent to each other (Robinson et al., 2005). The aim of an equivalence test was to show that the population means were equivalent to each other while the aim of a significance test was to show that there was a statistically significant difference between the population means if the null hypothesis is true (Lakens, 2017).

3. A two-sample Kolmogorov-Smirnov test ($p<0.05$ alfa level) was carried out to compare the overall shape of probability distributions. The K-S test is a non-parametric test for differences in the shape of two sample distributions (Wellek, 2010).
4. The Wilcoxon rank sum test was also used as a non-parametric permutation test to compare density plots ($p<0.01$) (Bowman and Azzalini, 1997).

2.3.4. Graphics visualisation tools

Ggplot2 is an R package based on the grammar of graphics (Wickham, 2010) a package that is based on the concept of building all graphs using the same components. Other graphics visualisation tools used included Graphviz/Notepad++ (https://www2.graphviz.org/) (https://notepad-plus-plus.org/); Lucidgraph (https://www.lucidchart.com/); Inkscape (https://inkscape.org/en/); Dia (http://dia-installer.de/); bean plots (Kampstra, 2008), alluvial diagrams and parallel coordinates (https://www.rstudio.com/).

2.3.5. Mapping/filtering

Manual mapping of the food groups in GUI with those from NPNS was carried out using Microsoft Excel 2016 (v16.0) as detailed in Sections 5.2 and 2.6. Filters were applied and shallow natural language processing (NLP) used to complete the mapping. Shallow NLP can be broadly described as a semi-automatic processing of human language in the area of information retrieval (Indurkhya and Damerau, 2010). Manual mapping was also carried out for free sugar estimation following criteria based on a modified version of methodology described previously Louie et al. (2015) (Sections 2.6 and 7.2).

2.3.6. Association analysis

Association analysis is a useful methodology for discovering interesting relationships hidden in a dataset which can be represented as sets of frequent items or in the form of association rules (Williams, 2011). This technique is commonly used in domains varying from market basket data to bioinformatics, medical diagnostics and data mining. In this study, the technique was used to explore the combination of component food items in meal pattern analysis (Sections 6.2 and 2.6).

2.4. Early childhood dental problems: classification tree analyses at 9 months and 3 years of age

2.4.1. Data and variables

This section of the study used classification tree analysis (CTA) to explore a complex network of infant/child and primary caregiver (PCG) psychosocial and physical health variables and identify key parameters related to early childhood dental problems. Data used for this analysis was derived from the infants in the GUI study at 9 months (Wave 1) and when the children were 3 years old (Wave 2). The dependent variable in the analysis was a PCG reported dental problem. At 9 months information on reported experience of dental problems for which the infant was taken to see a health care worker was recorded as a dichotomous variable with a positive response indicating a dental problem. In Wave 2 of the study, when the child was 3 years old, the PCG was asked: 'Has child been to visit the dentist because of a problem with his/her teeth?'. Again, this was recorded as a dichotomous variable with a positive response indicating a dental problem.

Independent variables were chosen based on their relevance to child dental health.

2.4.1.1. Sociodemographic variables

Socioeconomic and demographic variables selected were ethnicity and highest education level of the PCG, family social class and equivalised household annual income (Williams et al., 2013). The gender (male/female) of the child and age and gender of the PCG was recorded.

2.4.1.2. Variables related to health

At 9 months of age health was assessed by PCG global ratings of infant general health and whether the infant was admitted to a hospital ward because of an illness or health problem. The global health rating of the infant was dichotomised as 'Very healthy' or 'Not very healthy' in a similar fashion to previous studies with preschool children (Wake et al., 2008). At 3 years of age health was assessed by PCG global ratings of child health and whether the child ever had an accident or injury requiring hospital treatment or admission.

At both time points PCG (global) health was assessed by self-rating on a 5-point Likert scale.

2.4.1.3. Psychosocial variables/Behavioural habits

At 9 months infant temperament was assessed by the 24-item Infant Characteristics Questionnaire-ICQ (Bates et al., 1979) covering four domains: Fussy-difficult, Unadaptable, Dull (or Subdued) and Unpredictable. Other questions relating to behavioural habits were whether or not the infant used a soother/dummy in the past week and whether the PCG ever woke the baby at night for a feed. PCG stress levels were indicated by the 18-item Parental Stress Scale (Berry and Jones, 1995) and PCG depression was assessed using the 8-item short form measure of the Centre for Epidemiological Studies Depression Scale (Radloff, 1977). The 'Quality of Attachment' (QoA) to the infant was measured using the QoA subscale from the Maternal Postnatal Attachment Scale (Condon and Corkindale, 1998).

At 3 years child behaviour and emotional development were assessed using the 25-item Strengths and Difficulties Questionnaire (SDQ) (Goodman, 1997). Child temperament was measured using a modified version of the Short Temperament Scale for Toddlers (Prior et al., 2000). Questions relating to child oral behavioural habits included frequency of tooth brushing and how often the child sucked a soother or finger/thumb. The child-PCG relationship was assessed using the Child-Parent Relationship Scale (Pianta CPR-S) (Berry and Jones, 1995). Again, when the child was 3 years of age, PCG depression was assessed using the 8-item short form measure of the Centre for Epidemiological Studies Depression Scale yielding a total depression score. PCGs were categorised as 'depressed' or 'not depressed'. PCGs were also asked whether they had been 'treated for depression, anxiety or nerves'.

2.4.2. Data Analysis

Classification trees were generated with PCG reported experience of a dental problem at 9 months or at 3 years of age as the target variable for each output using IBM SPSS Statistics (v. 20.0: SPSS, Chicago, IL) and the Chi-squared Automatic Interaction Detection (CHAID) algorithm (Kass, 1980). Classification tree analysis is a non-parametric technique which repeatedly partitions the sample into subgroups based on the relationship with the target variable ('dental problem'). The most significant predictor or independent variable is used to split the group and this process repeated until there are no statistically

significant differences remaining in the subgroups with respect to the independent variable. In the model, CHAID maximum tree depth was set at 5, default parent and child node size settings were selected (parent node=100, child node=50), and the Bonferroni-adjusted chi-square statistic was used to determine node splitting and merging at a significance level of <0.05. After growing trees of greater depths, it was decided to prune to a tree depth of 4 or 5 as higher levels resulted in p-values approaching p=0.05 or minority class size numbers at the child nodes were small (<5). A 10-fold cross-validation procedure was carried out and both datasets evaluated to see if boosting or resampling would improve the prediction accuracy. The model tree for each dataset was selected and saved as a training sample and a higher misclassification cost was specified for misclassifying those with a 'dental problem' as 'no dental problem' (0.95) than for misclassifying those with 'no dental problem' as having a 'dental problem' (0.05). The classification results without undersampling or boosting were evaluated for both datasets. Random undersampling was carried out which modified the class distribution by discarding the majority class (Yap et al., 2014). The minority class in both datasets was the positive instances of having 'a dental problem' and the negative response was the majority class. The CHAID algorithm handles missing values by defining a separate category and then deciding whether to merge this or keep it separate.

Following on, a series of binary logistical regression analyses (forward-wald) was conducted to compare findings with those generated by the classification tree output.

2.5. Weight status and dental problems at 9 months and 3 years of age: classification tree analysis

2.5.1. Data and variables

Data used for this analysis was derived from the infants in the GUI study at 9 months (Wave 1) and followed up when the children were 3 years old (Wave 2) providing the data file with 9,793 cases.

2.5.1.1. Anthropometric measurements

A standard (Leicester) portable stadiometer was used to measure height of PCGs and children. The weight of the children was recorded using digital scales (SECA 835). PCG weight was recorded using a flat mechanical scale (SECA 761, Hamburg, Germany). The heights and weight for the PCG were fed-forward from Wave 1 and were not retaken unless missing or noted for rechecking (Murray et al., 2013). The RMF files were used for all calculations of BMI as the BMI values for the AMF were calculated based on the top and bottom-coded weights and heights included in that datafile rather than the original measurements that were included in the RMF.

BMI was calculated as weight divided by height squared (kg/m^2) and, for children, classified as overweight, obese, normal weight or thinness according to the IOTF age and gender specific cut-offs for 3-year olds (Cole et al., 2000, Cole et al., 2007).

Table 2.2 International Obesity Task Force (IOTF) Body Mass Index Cut-Offs for Thinness, Overweight and Obesity in 3 year old Children.

BMI classification	BMI Kg/m^2 cut-off point at age 18	Male 3 years of age	Female 3 years of age
Thinness	<18.5	<14.83	<14.60
Normal range	18.5-24.99	14.83-17.849	14.6-17.639
Overweight	≥25	17.85-19.499	14.64-19.379
Obese	≥30	19.5-20.749	19.38-20.739
Morbidly obese	≥35	≥20.75	≥20.74

Source: Amended from Cole and Lobstein (2012).

For simplicity, the classification of thinness (low BMI for age) was also described as underweight in this analysis although the latter strictly means low weight for age in children. It is common practice to report weight status using two distinct methods of estimation. Thus, overweight and obesity for children was also classified using the UK adaptation of percentile cut-offs from the WHO Multicentre Growth Reference Study (MGRS) with overweight criteria defined

as a BMI between the 91st and 98th percentile while obesity was defined as a BMI on or greater than the 98th percentile (Cole et al., 2012). LMS parameters were created using Cole's LMS method (Flegal and Cole, 2013, Cole, 1990) and the SPSS syntax used included in Appendix A. PCG BMI was categorised into underweight (BMI<18.5), normal (BMI 18.5-24.9), overweight (BMI 25-29.9) and obese (BMI>30).

2.5.1.2. CTA target variable

The dichotomous target variable was a PCG reported dental problem. The question asked was: Has <child> been to visit the dentist because of a problem with his/her teeth?

2.5.1.3. CTA predictor variables

Attributes (independent variables) that were relevant to the target variable (dependent variable) were selected for inclusion in the model based on findings from Chapter 3. The demographic and socioeconomic variables selected were child gender, PCG age and gender, ethnicity, PCG education level, family social class and annual equivalised household income (Gussy et al., 2006, Harris et al., 2004, Hayden et al., 2012, Hooley et al., 2012c, van der Tas et al., 2016, Hooley et al., 2012a). Ethnicity was defined as Irish, Any other White background, Black, Asian or Other. The highest education level attained by the PCG was one of thirteen categories ranging from no formal education to doctorate level which was collapsed to five groups for descriptive analysis. Family social class was measured using the Irish Central Statistics Office's classification based on occupation, categorising families into one of seven groups which was collapsed to four groups for descriptive analysis. Annual disposable household income was calculated by using an equivalence scale to "weight" each household for differences in size and composition with respect to number of adults and children (Murray et al., 2013). Markers of health status (Sheiham, 2006, Harris et al., 2004, Sheiham and Watt, 2000, Kantovitz et al., 2006) included PCG reported child illness, disability, allergies and injuries, as well as TV-viewing hours, tooth-brushing, soother/thumb-sucking, and breastfeeding (duration and if ceased now) as markers of health behaviour (Harris et al., 2004, Hooley et al., 2012a, Chaffee et al., 2015, Layte et al., 2014b). Dietary intake (Marshall et al., 2007b, Chaffee et al., 2015, Harris et al., 2004) was assessed using a modified version of the Sallis-Amherst Food Frequency Questionnaire from the Longitudinal Study of Australian Children (LSAC) (Sallis et al., 2002). PCG reported the child's frequency of consumption

of 15 food categories (e.g. sweets, fizzy drinks/minerals/cordials) over the previous 24 hours as once, more than once or none at all.

2.5.2. Data analysis

Wave 1 GUI data were statistically re-weighted to represent the population. Wave 2 data was weighted for attrition between waves and emigration combined with the Wave 1 weight (Quail et al., 2011). For this analysis, the following parameters were selected in either SPSS (v. 20.0: SPSS, Chicago, IL) or SPSS Modeler (IBM SPSS Modeler v. 14.2: Chicago, IL) using the Chi-squared Automatic Interaction Detection (CHAID) algorithm (Kass, 1980): maximum tree depth=5, parent node=100, child node=50 and Bonferroni-adjusted chi-square statistic, significance <0.05. A 10-fold cross-validation assessed model performance and produced an average misclassification risk. To compensate for class imbalance, a higher misclassification cost was specified for misclassifying those with a 'dental problem' as 'no dental problem' (0.95) than for misclassifying those with 'no dental problem' as having a 'dental problem' (0.05). The degree of missing cases in the 3-year old GUI infant cohort was small except for the PCG BMI (5.2%), equivalised annual income (5.5%) and child BMI (2.6%), as previously reported (Layte et al., 2014b). The CHAID algorithm handled missing values by defining a separate category and treating them as a single category so that they are not excluded in the analysis (Maimon and Rokach, 2005). A binary logistical regression analysis (forward-wald) was also conducted to compare findings with those generated by the classification tree output. A confusion matrix for a binary classifier provided an estimation of selected performance metrics.

2.6. Measuring dietary intake

2.6.1. Data collection and participants

This research utilised data collected as part of two studies: the second wave of the GUI infant cohort longitudinal survey which was carried out by the joint Economic Social Research Institute-Trinity College Dublin (ESRI-TCD) GUI study team from December 2010 to July 2011 and the NPNS cross-sectional study which was conducted by Irish Universities Nutrition Alliance (IUNA) from October 2010 to September 2011. The second wave of the GUI infant cohort were 3 years of age at the time of interview (n= 9,793). The NPNS had a total

sample of 500 children aged 2-4 years; but only the 3-year olds were included for this analysis (n=126). Both samples were nationally representative, and surveys were conducted at a similar time.

2.6.2. Food intake measurement

In the GUI study, dietary intake was assessed using a SFQ, previously used in the Longitudinal Study of Australian Children (LSAC), to characterise healthy and unhealthy food intake (Sallis et al., 2002). The PCG reported how frequently their child consumed 15 food categories during the previous 24 hours. Intakes were recorded as once, more than once, or none at all. No information on food portion size was recorded. Foods were categorised as "healthy" or "unhealthy" as follows: fresh fruit; cooked vegetables; raw vegetables or salad; hamburger, hot dog, sausage, meat pie; hot chips or french fries; crisps or savour snacks; biscuits, doughnuts, cake, pie or chocolate; sweets; full fat cheese/yoghurt/fromage frais; low fat cheese/low fat yoghurt; water (tap, still sparkling); fizzy drinks/minerals/cordial/squash (diet); fizzy drinks/minerals/cordials/squash (not diet); full cream milk or full cream milk products; skimmed/semi-skimmed milk or milk products. The survey questionnaire containing the SFQ is available at: http://www.ucd.ie/issda/data/growingupinirelandgui/.

A 4-day weighed food record was used in NPNS to collect food and beverage intake data (Irish Universities Nutrition Alliance, 2012). At least one of the 4 days included a weekend day and a nutrition researcher trained caregivers on how to use the food diary and weighing scales to record intakes. The caregivers were requested to record information relating to the amount, brand and type of foods and beverages consumed by the child and to include cooking method, recipes, packaging type, food leftover and time of eating occasion. Food and beverage intake data were reported, after weighing, in grams. The quantification protocol used is available at http://www.iuna.net/ and has been previously reported (Walton and Flynn, 2013). In total, there were 1,652 different food codes in the NPNS and each food was also assigned to one of 77 food group categories. Data from McCance and Widdowson (2002) were used to estimate nutrient intake using WISP© (Tinuviel Software, Anglesey, UK) as described elsewhere (Irish Universities Nutrition Alliance, 2012). Data analysed did not exclude under-reporters which were estimated previously for this cohort (Walton, 2012, Walton et al., 2017).

2.6.3. Data preparation and mapping protocol

Data files were imported from SPSS (v. 20.0: SPSS, Chicago, IL) or converted to .csv format before importing to R (version 3.2.2) for linkage and analysis (Further details in Appendix). The 77 Food group categories were used for this analysis and other variables such as food name, cooking method, day of consumption, meal-type and food description were also selected. A unidirectional mapping procedure (Figure 2.3) was carried out using a shallow natural language processing (NLP) approach.

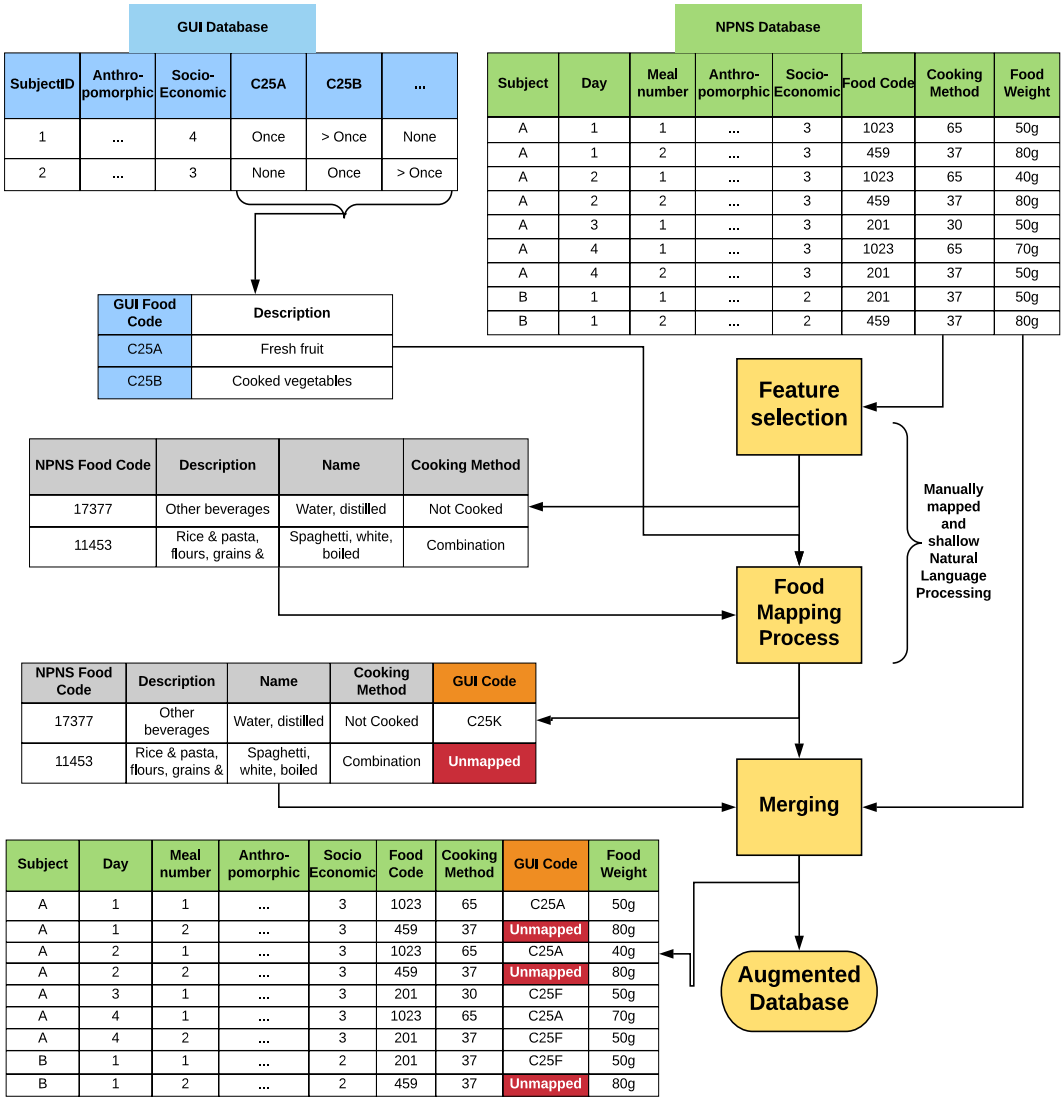
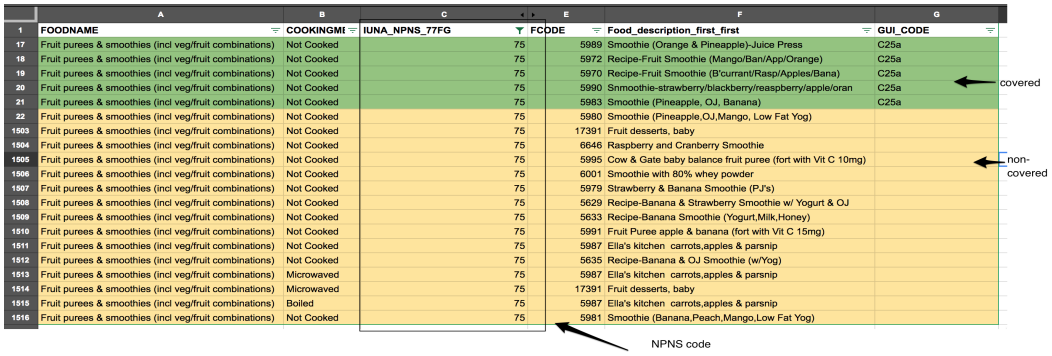


Figure 2.3 Flow diagram showing data processing steps for unidirectional mapping of GUI food codes with NPNS food codes. GUI: Growing Up in Ireland; NPNS: National Preschool Nutrition Survey. Feature selection identified variables from both GUI and NPNS databases that were desired, e.g. socioeconomic class, cooking method, food weight.

All food categories in NPNS were sorted, grouped and filtered to facilitate easy mapping whereby all GUI food groups were filled with information from the NPNS food datafile and consolidated into a single augmented database. The augmented data were analysed to examine all food groups described in NPNS and GUI and what proportion of foods were *covered*, *non-covered* or *partially-covered* by GUI food groups relative to the NPNS database which included a more detailed dietary record. The term *non-covered* indicated a specific food consumption that could not be mapped using a GUI food code, i.e., the food in NPNS is not matched by the same food in GUI. If a food category in NPNS was only *partially covered* by a GUI food description, then the resulting mapped GUI database was only *partially-covered* for that food category.



	A	B	C	E	F	G
1	FOODNAME	COOKINGMI	IUNA_NPNS_77FG	FCODE	Food_description_first_first	GUI_CODE
17	Fruit purees & smoothies (incl veg/fruit combinations)	Not Cooked	75	5989	Smoothie (Orange & Pineapple)-Juice Press	C25a
18	Fruit purees & smoothies (incl veg/fruit combinations)	Not Cooked	75	5972	Recipe-Fruit Smoothie (Mango/Ban/App/Orange)	C25a
19	Fruit purees & smoothies (incl veg/fruit combinations)	Not Cooked	75	5970	Recipe-Fruit Smoothie (B'currant/Rasp/Apples/Bana)	C25a
20	Fruit purees & smoothies (incl veg/fruit combinations)	Not Cooked	75	5990	Smmoothie-strawberry/blackberry/reaspberry/apple/oran	C25a
21	Fruit purees & smoothies (incl veg/fruit combinations)	Not Cooked	75	5983	Smoothie (Pineapple, OJ, Banana)	C25a
22	Fruit purees & smoothies (incl veg/fruit combinations)	Not Cooked	75	5980	Smoothie (Pineapple,OJ,Mango, Low Fat Yog)	
1503	Fruit purees & smoothies (incl veg/fruit combinations)	Not Cooked	75	17391	Fruit desserts, baby	
1504	Fruit purees & smoothies (incl veg/fruit combinations)	Not Cooked	75	6646	Raspberry and Cranberry Smoothie	
1505	Fruit purees & smoothies (incl veg/fruit combinations)	Not Cooked	75	5995	Cow & Gate baby balance fruit puree (fort with Vit C 10mg)	
1506	Fruit purees & smoothies (incl veg/fruit combinations)	Not Cooked	75	6001	Smoothie with 80% whey powder	
1507	Fruit purees & smoothies (incl veg/fruit combinations)	Not Cooked	75	5979	Strawberry & Banana Smoothie (PJ's)	
1508	Fruit purees & smoothies (incl veg/fruit combinations)	Not Cooked	75	5629	Recipe-Banana & Strawberry Smoothie w/ Yogurt & OJ	
1509	Fruit purees & smoothies (incl veg/fruit combinations)	Not Cooked	75	5633	Recipe-Banana Smoothie (Yogurt,Milk,Honey)	
1510	Fruit purees & smoothies (incl veg/fruit combinations)	Not Cooked	75	5991	Fruit Puree apple & banana (fort with Vit C 15mg)	
1511	Fruit purees & smoothies (incl veg/fruit combinations)	Not Cooked	75	5987	Ella's kitchen carrots,apples & parsnip	
1512	Fruit purees & smoothies (incl veg/fruit combinations)	Not Cooked	75	5635	Recipe-Banana & OJ Smoothie (w/Yog)	
1513	Fruit purees & smoothies (incl veg/fruit combinations)	Microwaved	75	5987	Ella's kitchen carrots,apples & parsnip	
1514	Fruit purees & smoothies (incl veg/fruit combinations)	Microwaved	75	17391	Fruit desserts, baby	
1515	Fruit purees & smoothies (incl veg/fruit combinations)	Boiled	75	5987	Ella's kitchen carrots,apples & parsnip	
1516	Fruit purees & smoothies (incl veg/fruit combinations)	Not Cooked	75	5981	Smoothie (Banana,Peach,Mango,Low Fat Yog)	

Figure 2.4 Code snippet illustrating manual mapping of GUI food codes and an example of a *partially-covered* group (Fruit purees and smoothies), showing the food name, cooking method, NPNS food group code (n=77), NPNS individual food code, food description and GUI food code. Food descriptions in area shaded green is *covered* while those in area shaded orange is *non-covered*. Food consumption was described as *covered* if there was a matching GUI food group that the food consumption could be fully mapped to, i.e., the food in NPNS was matched by the same food group in GUI.

For example, Figure 2.4 illustrates a snippet from the manual mapping for various categories of Fruit purees and smoothies. Some of these were mapped to GUI food code descriptions for 'Fresh fruit and vegetables' (top rows, shaded dark green, labelled *covered* and GUI code C25a) while others were not (orange shaded area labelled *non-covered*) as they were deemed not to constitute only fresh fruit and vegetables. Thus, this GUI food code group would result in a *partially-covered* group.

Subj ID	Survey Day	Meal Type	Meal No	Time	FWT	Food description	NPNS_ 77FG	GUI_ CODE
108	1	1	1	07:45	30	Recipe-Brown Soda Bread	4	NA
108	1	1	1	07:45	122	Porridge made with whole milk (Irish Recipe)	7	NA
108	1	1	1	07:45	3	Fat spreads (59% Fat) Not polyunsaturated (Irish)	22	NA
108	1	1	1	07:45	4	Jam, fruit with edible seeds	57	NA
108	1	1	1	07:45	20	Water, distilled	65	C25k
108	1	2	2	12:00	17	Bunalun Yogurt/Milk Chocolate Rice Cakes (not fort)	8	C25g
108	1	2	2	12:00	21	Golden Vale Cheese Strings	14	C25i
108	1	2	2	12:00	100	Fortified food-Drinking Yoghurt-Actimel (Danone)	15	C25i
108	1	2	2	12:00	90	Kid's Yoghurts (Irish) (13-14% sugars)	15	C25i
108	1	2	2	12:00	55	Orange juice, unsweetened	35	NA
108	1	2	2	12:00	176	Bananas, weighed with skin	36	C25a
108	1	2	2	12:00	50	Grapes, average	37	C25a
108	1	2	2	12:00	14	Raisins	37	C25a
108	1	2	2	12:00	88	Water, distilled	65	C25k
108	1	7	3	16:00	17	Bunalun Yogurt/Milk Chocolate Rice Cakes (not fort)	8	C25g
108	1	7	3	16:00	11	Sandwich biscuits, cream filled	8	C25g
108	1	5	4	17:00	3	Ice cream wafers	8	C25g
108	1	5	4	17:00	28	Sponge cake, with dairy cream and jam	9	C25g
108	1	5	4	17:00	34	Ice cream, dairy, vanilla	16	NA
108	1	5	4	17:00	31	New potatoes, boiled in unsalted water	25	NA
108	1	5	4	17:00	24	Peas, frozen, boiled in unsalted water	29	C25b
108	1	5	4	17:00	57	Orange juice, unsweetened	35	NA
108	1	5	4	17:00	30	Ham, gammon joint, boiled	43	NA

Table 2.3 Food consumption entries for day 1 of survey for child subject ID 108, depicting survey day, meal type, meal number, time, food weight (FWT), food description, NPNS food code, GUI food code where mapped and NA indicating *non-covered*.

A consumption in NPNS was defined as any eating occasion (EO) of a food or drink (snack or main meal) and an entry in the food diary was considered a consumption as indicated in Table 2.3 which shows the food entries for child subject number 108.

2.6.4. Aligning two surveys- GUI and NPNS

To determine if the subjects in NPNS and GUI were sampled from the same population it was necessary to compare the estimates of food intake consumption from both surveys. If both surveys were aligned, then food consumption data from NPNS could be used directly to report on the GUI sample population. Following the mapping procedures, a series of proportion tests were carried out to assess if the GUI and NPNS surveys could be aligned. The assumption proposed was that both datasets were sampled from the same population. The protocol is outlined in Figure 2.5.

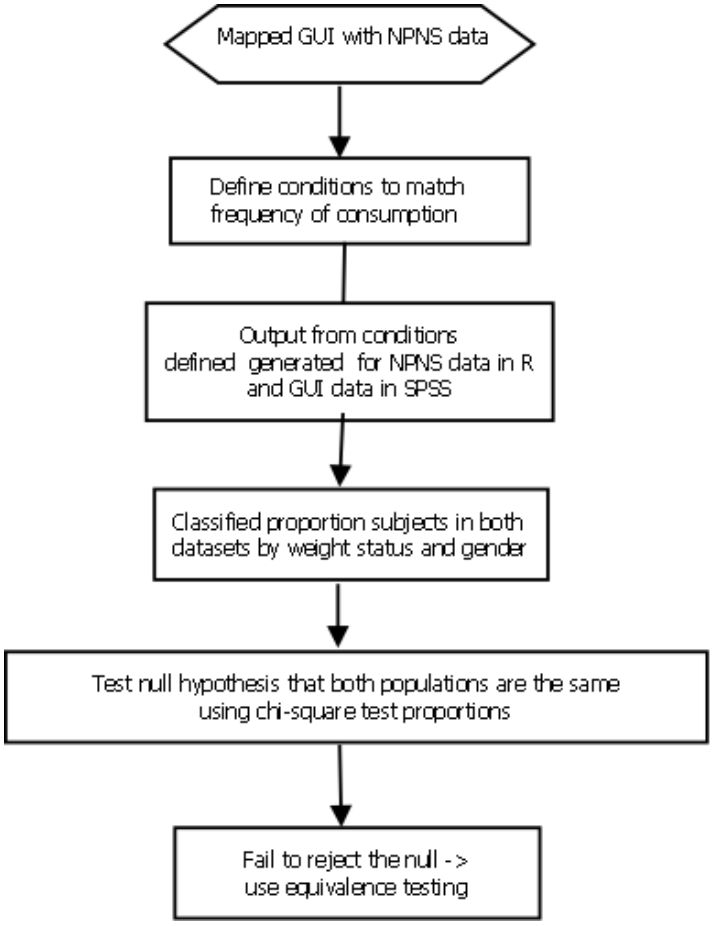


Figure 2.5 Protocol for aligning GUI and NPNS surveys.

Initially, the conditions for defining frequency of consumption were used to generate the proportion of subjects in each category for NPNS according to the GUI food code mapped. The proportion of subjects in GUI for each category of weight and gender was also estimated according to the GUI food code. The Z score for each proportion was calculated and a p-value estimated according to the chi-square test for proportion. A series of equivalence tests were also calculated. This was repeated for each of the three different defined sets of conditions as outlined below.

As the two surveys used different instruments to estimate food intake it was necessary to propose a set of conditions to “match” frequency reporting. The GUI SFQ recorded responses to food group consumption in the previous 24 hours as ‘none’ (zero), ‘once’ or more than once. The objective was to align the 1 day SFQ to the 4 day FD. This procedure started by firstly defining ‘zero’ as the absence of consuming a specific food. Then the definition of ‘zero’ was relaxed, meaning that if the food was consumed less than n times over the 4

days the frequency for the food was set to zero. Gender and weight status using the IOTF classification were taken into consideration when computing the frequencies. The approaches to define “zero” consumption considered were:

1. “Total frequency zero”: no consumption of the food during the 4-day period of NPNS
2. “Total frequency one”: food consumed once during the 4-day period of NPNS
3. “Single day multiple consumption”: this included the previous definitions plus those foods that were consumed multiple times in only one day of the NPSN survey

For each definition of zero above the percentage of GUI and NPNS subjects in each weight category (using IOTF criteria) was calculated for each food code. The chi-square test for homogeneity (or two proportions) was used to test the null hypothesis ($p<0.05$, alfa level) that there was no difference between the two sample populations under the conditions of each of the approaches defined above.

A further series of approaches were used to define “once” and “more than once” based on the “zero” definitions already tested. For example, defining zero consumption as “total frequency one”, a food consumption frequency of “once” in GUI was defined as a maximum average of 1.0 over the 4-day NPSN period. The chi-square test for proportions was repeated for all these categories in GUI and NPNS.

Finally, a series of equivalence tests were carried out to test if the populations were the same (Robinson et al., 2005).

2.6.5. Quantitative analysis of mapped data and augmented database

The initial aggregation was done at the subject and survey day levels meaning that for each subject and each day of the survey an aggregated record was obtained. Aggregate metrics were defined and determined for all food items included mean, interquartile range, maximum, minimum, standard deviation and standard error of the mean. Aggregates estimated included the frequency and amount (g/day) of *covered*, *non-covered* and *partially-covered* food groups which were also expressed as a percentage of the total amount of food consumed. The analysis treated each day of the 4 days in NPNS as an independent day. The mean daily intake amount and the frequency of each

food consumed was calculated for each NPNS participant. The amount and frequency of food consumed was also calculated using an average consumption method (Section 2.6.6).

The total number of times when a *non-covered* food was consumed (total consumption frequency per day) and the total food amount (g/d) of a *non-covered* food was calculated. The ratio of the frequency of consumption of *non-covered* food over the total food frequency was determined. A similar ratio was determined for the amount of *non-covered* food consumed over amount of total food consumed. *Partially-covered* food groups were also included in the analyses and included in either the *covered* or *non-covered* category after the final mapping. The frequency distributions of the ratio of consumption frequency and amount of *non-covered* food consumed divided by the total food consumed were displayed as histograms. Using a non-parametric density estimation, the distribution of the proportion of *non-covered* food consumed each day of the week was displayed graphically using kernel density plots and tested formally using the Wilcoxon rank sum test ($p<0.01$) (Dalgaard, 2008).

2.6.6. Data preparation and analysis for cariogenic food intake and meal analysis

The pattern of consumption of selected cariogenic food and drink (CF) in 3 year old preschool children was investigated using the NPNS food database mapped with food intake codes for 3 year olds from the GUI survey (section 2.6.3). This mapped dataset was then analysed to explore the pattern of dietary intake of selected CF items as snacks or main meals. The mean daily intake amount (g/day) and the frequency of each food consumed was calculated for each NPNS participant, by summing the amount of all foods a subject consumed per food code, averaging across the four days for each subject and then calculating the total sample average. Frequency was estimated by summing the total number of times the food appeared in the diary and dividing by four, i.e. the number of days in the survey. The average consumption amount (g/day) and frequency for each food was computed by aggregating the data across each subject and each survey day before averaging the food weight. Thus, if a food was consumed more than once on a given day the average consumption amount was calculated to provide a closer representation of the actual amount of food consumed on a single EO. If a food was not consumed at all on a given day, it was not included in the estimations for average consumption amount. Therefore, the average consumption gave an

estimate of portion size per EO, rather than the average amount consumed over the 4-day record or mean daily intake. For most food items the average consumption is generally greater than the mean daily intake of that food. All results reported were for consumers only. Analysing intake using average consumption estimates and consumer-only data provided the "worst-case" exposure as it prevented reduction of the estimates through inclusion of days or individuals with no intake of a particular cariogenic food or drink (CF) (Connolly et al., 2010).

Global statistics were generated for the number of subjects who either never consumed a CF or consumed a CF. Each meal type (both 'snacks' and 'meals') was defined by its food components. A food component was defined as a single food item from each meal. The number of food components, both CF and non-cariogenic food and drink (NCF), in a meal, were determined at the meal level and at the subject level.

Bean plots were generated in R for dietary intake estimated using both mean daily intake and average consumption methods. These plots are an "alternative to the boxplot for visual comparison of univariate data between groups" (Kampstra, 2008) as outlined in Section 1.7.4. The bean plot uses a combination of non-parametric kernel density estimates of the probability density function with a scatter plot of all data points. A kernel density estimates the probability density function of a random variable in a non-parametric fashion (Williams, 2011). The individual observations are depicted as short horizontal lines in a one-dimensional scatter plot and the estimated density shape of the distributions is displayed as a polygon. The name derives from green beans as the density shape represents the pod and the scatter plot shows the seeds inside the pod. The heavy solid horizontal line represents the average for each subgroup and the dashed horizontal line is the overall average value for the dataset. Asymmetric bean plots allow for easy comparison between two subgroups such as main meals and snacks and reveal anomalies such as bimodal distributions (Kampstra, 2008). Main meals were coloured blue while snacks were coloured red. CF items (rice puddings and custard, tinned fruit and carbonated beverages) with a small number of subjects (<20) were not included as the number of observations were too low to generate a meaningful plot. A two-sample Kolmogorov-Smirnov test ($p<0.05$ alfa level) was carried out to determine if the distributions of CF items as a snack or main meal, as estimated by both mean daily intake and average consumption, were similar.

2.6.7. Meal and CF pattern analysis

Association analysis was carried out using the NPNS database mapped with the GUI food codes to: (1) identify the food components that comprised a meal; (2) compare meals by chaining the components within a meal using either the GUI coding or the NPNS 77 Food group coding; (3) identify the combinations of food components that characterise the most frequently consumed meals and (4) assess the dietary interaction of each CF with NCF and the other CF items selected. Variables included were: subject ID, meal type, time of consumption, NPNS 77 FG, food weight and GUI food code. An association analysis algorithm generated frequent item sets (Williams, 2011). Association analysis is a useful methodology for discovering interesting relationships hidden in a dataset. The algorithm iterated through each subject and day of the survey and identified each meal that the subject consumed by sorting the records in the diary by time of the day and meal type. For each subject, all food and drink consumed at the same time of the day and coded with the same meal type were considered part of a distinct meal. After identification of each meal the data was restructured into a tree shape where each branch was a subject and each leaf was a list of days and corresponding meals. Two distinct analyses were then carried out by going through each of the leaves of the tree and retaining the metadata (subject information) of each branch. The first analysis used a set of distinct GUI food codes and labelled foods *not-covered* by GUI as *non-covered* to build a unique identifier for each meal. This was called key and this meal identifier was the key for GUI-derived food codes. A key-value pair is a fundamental data representation in computing systems and applications and is also described as an attribute-value pair (Williams, 2011). The second analysis repeated this process but used a set of distinct NPNS 77 Food group codes and was defined as key for NPNS-derived food codes. Finally, the association tree was re-shaped using “cariogenic” and “non-cariogenic” descriptors to describe the combinations and interactions between CF consumed as snacks or a main meal.

2.7. Sugar intake and free sugar mapping

Figure 2.5. outlines the algorithm used to carry out the FS mapping which was based on a modified version of the method developed by Louie et al (Louie et

al., 2015) to estimate added sugar (AS does not include sugars naturally present in honey, syrups and unsweetened fruit juices).

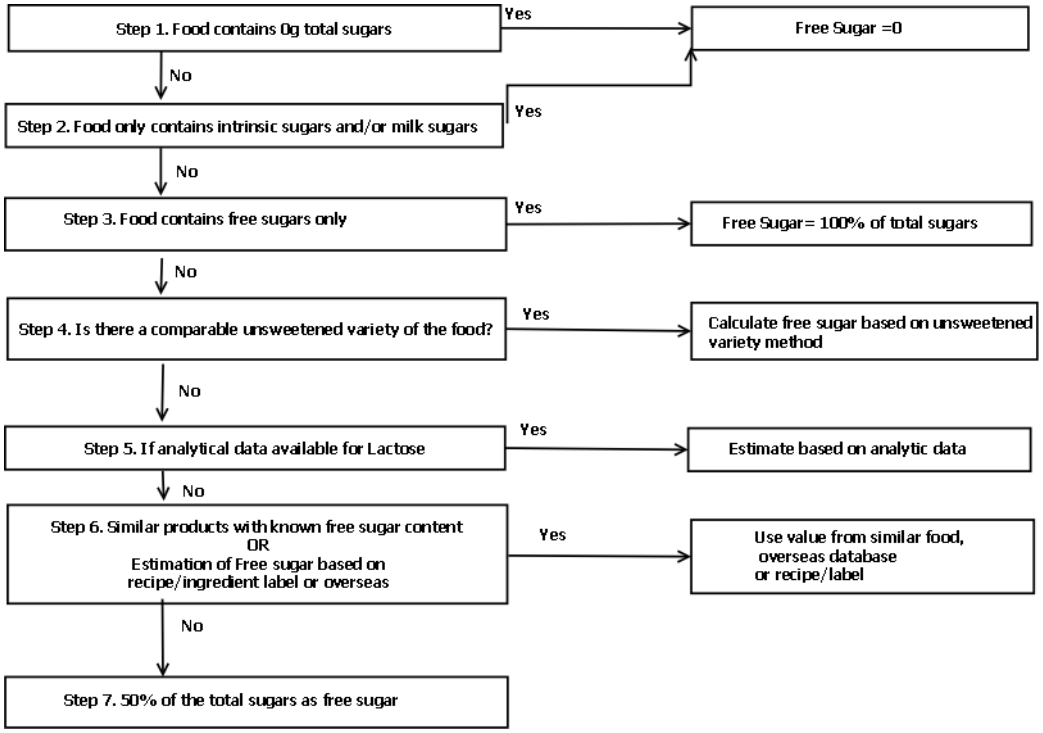


Figure 2.6 Decision algorithm for estimating free sugars content of national preschool nutrition survey (NPNS) foods. Amended from Louie et al. (2015).

Briefly, the steps were followed starting from step 1 until the food fulfilled the criteria for assignment at a step. If a food did not meet the criteria in any of steps 1-6 then FS was set at 50% total sugar (TS) (step 7). To carry out the FS mapping, foods were assigned 0 g of FS if steps 1 or 2 applied. All the total sugar (TS) was assigned as FS for foods at step 3. If an unsweetened variety of the same food existed, then FS estimation was based on the difference between the sweetened and unsweetened variety at step 4. At step 5 if analytical data existed for lactose or other sugars then an estimate of FS was based on subtracting lactose. The more subjective steps were 6 and 7. These involved a subjective decision to use a recipe, ingredient label or overseas database such as the AUStralian Food and NUTrient Database (AUSNUT) 2011-13 food composition database (Food Standards Australia New Zealand, 2014). Step 7 which assigned 50% of TS as FS was only used if there was no other means of estimating FS content.

The TS and lactose estimates from the NPNS database were imported into RStudio and then exported to a .csv file before manually estimating FS content

based on the rules described above. Two researchers (AOS and MC) estimated FS in the NPNS food database. The results were compared and where differences occurred in FS estimation the final value was based on agreed interpretation or an average between the two estimates. The mapped GUI database (Section 2.6.3) was then augmented by importing the FS estimates (Appendix A). This FS mapping was also compared to a previously reported estimation of FS using the same cohort of 3-year olds (Newens and Walton, 2016) and the distributions compared using the Kolmogorov-Smirnov tests ($p<0.01$). Quantitative analysis and metrics similar to that detailed in chapter 2.6.5 were carried out. All statistical analyses were carried out using R Studio (https://www.rstudio.com/).

After the data was loaded into RStudio it was aggregated across:

1. Subject ID
2. Day of the week
3. Day of the survey

To compare how well the GUI-SFQ captured TS and FS intakes the following metrics were computed:

1. The total number of times when a non-GUI consumption occurred.
2. The total food weight for non-GUI consumptions.
3. The total sugars (weight) for a non-GUI consumption.
4. The total sugars for a GUI consumption.
5. The total number of times when a GUI consumption occurs.
6. The total sugars.
7. The total number of consumptions.
8. The total food weight.

The following ratios were then derived for both TS and FS:

1. Ratio of the count (frequency) of non-GUI over total food
2. Ratio of the food weight of non-GUI over total food weight.
3. Ratio of non-GUI sugars over the total sugars
4. Ratio of GUI sugars over the total sugars.

Complete details of these analyses are reported in the code documents detailed in Appendix A. TS and FS were determined by multiplying the weight of food consumed daily, aggregated at the subject level, by the percentage of TS or FS. The mean daily intake of TS and FS (g/d), frequency of consumption and as a percentage of total energy intake, (TEI) were presented as summary

statistics. The percentage of consumers of each food group were calculated. The probability of consuming a food or drink as a snack or main meal was estimated by using the total count of snacks or main meals over all 4 days of the survey. The daily intake of TS and FS *covered* and *non-covered* by GUI food groups by amount (g/day) and as a percentage of total energy intake (%TEI) were presented as bar graphs. In NPNS there were 1,652 food codes which were categorised into 77 food groups and further re-categorised into 19 food groups. In GUI, there were 15 food groups. Food groups for both GUI and NPNS were also recategorised as follows to highlight the main FS food sources: bread and cereals, RTEBC, cakes and biscuits, dairy products, desserts and puddings, fruit and vegetables, Fruit juice and smoothies, sugar and syrups, chocolate confectionary, non-chocolate confectionary, soft drinks (non-diet), soft drinks (diet) and other. Dairy products included all milk, yoghurt, cheese and ice-cream products. Soft drinks (non-diet) included carbonated beverages, squashes, cordials and fruit juice drinks. Breads and cereals included all rice, pasta, grains and cereal based products except RTEBC. Full details of the mapped food groups are listed in Appendix A. Sugar intake (TS and FS) as a percentage of TEI was calculated using 0.017 MJg-1 ((World Health Organization, 2003). The percentage of the sample population with a FS intake greater than the WHO recommendations (World Health Organization, 2015) were determined.

Chapter 3. Early childhood dental problems: classification tree analyses of two waves of an infant cohort study

3.1. Introduction

The prevalence of oral health problems in young children has increased in recent years, following a decline in previous decades (Bourgeois and Llodra, 2014, Dye et al., 2010). The bidirectional relationship between oral health problems and child health and development is complicated by a variety of sociodemographic influences (Hooley et al., 2012b, Sheiham, 2006). In addition, the primary caregiver (PCG) is the gate-keeper in providing and promoting general and oral health care for the developing child. Therefore, PCG health and wellbeing is inextricably linked to child health and ultimately defined by similar social determinants (Moimaz et al., 2014).

In recent years, dental research has expanded, recognising that psychosocial, behavioural and environmental factors significantly impact oral health outcomes (Fisher-Owens et al., 2007, Newton and Bower, 2005). A number of studies have reported relationships between PCG psychological distress, child socio-emotional behaviour or infant temperament and child oral health outcomes (Tang et al., 2005, Menon et al., 2013, Quinonez et al., 2001a, Spitz et al., 2006, Aminabadi et al., 2014). Parental stress and depression may impact on the caregivers' ability to impart preventive oral health measures at critical developmental stages (Tang et al., 2005, LaValle et al., 2000) and are often related to aspects of infant temperament and child socio-emotional behaviour (Spitz et al., 2006, Renzaho and Silva-Sanigorski, 2013, Mäntymaa et al., 2006). Depressive symptoms in mothers may lead to inconsistent parenting and unhealthy feeding habits (Kim Seow, 2012). Furthermore, a positive child temperament appears to be protective against early childhood caries (ECC) while a difficult temperament and poor feeding practices are both equally strongly associated with ECC (Aminabadi et al., 2014). Influenced by

behavioural and social science research, conceptual models of oral health have developed to incorporate a wider framework, including psychosocial and behavioural factors which expand beyond the individual, to the family and community level (Kim Seow, 2012, Fisher-Owens et al., 2007). While several studies have focused on the socio-demographic components within models, few have examined the role of psychosocial and behavioural factors in large population studies (Hooley et al., 2012b).

Classification and regression trees have been used in clinical settings for risk assessment or diagnostic prediction but less so in public health research (Kuhn et al., 2014). The aim of classification tree analysis (CTA) is to create a model that predicts a target outcome (dependent variable) based on the strength of interactions between categorical or continuous input variables (independent variables) (Loh, 2014). To date, most of the research on early childhood dental problems and the health and psychosocial attributes of the child and PCG have concentrated on the effect of a single variable using relatively small sample sizes (LaValle et al., 2000, Hooley et al., 2012b). This study uses CTA to explore a complex network of infant/child and PCG psychosocial and physical health variables and identify key parameters related to early childhood dental problems in a large nationally representative sample of Irish children in infancy and again in early childhood.

3.2. Methods

3.2.1. Data and variables

This section of the study used CTA to explore a complex network of infant/child and PCG psychosocial and physical health variables and identify key parameters related to early childhood dental problems. Data used for this analysis was derived from the infants in the GUI study at 9 months (Wave 1) and when the children were 3-years old (Wave 2). The dependent variable in the analysis was a PCG reported dental problem.

Independent variables were chosen based on their relevance to child dental health. Socioeconomic and demographic variables selected were ethnicity and highest education level of the PCG, family social class and equivalised household annual income, child gender and age and gender of the PCG. Health was assessed by PCG global ratings of general health and whether the infant

or child required hospital admission or treatment. Other variables included at 9 months of age were infant temperament and questions relating to behavioural habits such as whether or not the infant used a soother/dummy in the past week and whether the PCG ever woke the baby at night for a feed. PCG stress levels, depression and the 'Quality of Attachment' (QoA) to the infant were also measured. At 3 years child behaviour and emotional development were assessed using the Strengths and Difficulties Questionnaire. Child temperament was measured using a modified version of the Short Temperament Scale for Toddlers. Questions relating to child oral behavioural habits included frequency of tooth brushing and how often the child sucked a soother or finger/thumb. The child-PCG relationship was assessed using the Child-Parent Relationship Scale. Again, when the child was 3 years of age PCG depression was assessed.

3.2.2. Data Analysis

Classification trees were generated with PCG reported experience of a dental problem at 9 months or at 3 years of age as the target variable for each output using IBM SPSS statistics (v. 20.0: SPSS, Chicago, IL) and the Chi-squared Automatic Interaction Detection (CHAID) algorithm (Kass, 1980). The Bonferroni-adjusted chi-square statistic was used to determine node splitting and merging at a significance level of <0.05. Following on, a series of binary logistical regression analyses (forward-wald) was conducted to compare findings with those generated by the classification tree output. Further details are included in Section 2.4.

3.3. Results

3.3.1. Profile of the sample

At 9 months of age 2.7% (n=302) of infants had a PCG-reported dental problem for which they had sought care from a health care professional. When children were 3 years old 5.0% (n=493) had a dental problem for which they had sought care from a dentist (Table 3.1). The sociodemographic profile of the study populations was similar at both time points apart from the variable reflecting family income (Table 3.1). The mean equivalised annual household income was lower at Wave 2 compared to Wave 1.

Table 3.1 Demographics and profile of Growing Up in Ireland Infant Cohort Study Participants.

Characteristic	Total at 9 months old	Total at 3 years old
Survey Date	Sept 2008/April	Dec 2010/June
Sample Size, n	11,134	9,793
Child gender, n (%)		
Boy	5715 (51.3)	5024
Girl	5419 (48.7)	4769
Dental problem, n (%)	302 (2.7)	493 (5.0)
Hospital admission (ever)*, n (%)		
Yes	1453 (13.1)	1569
No	9674 (86.9)	8201
Global health child, n (%)		
Very healthy	9197 (82.6)	7312
Not very healthy	1895 (17.0)	2476
Infant Characteristics Questionnaire, mean (SD)		
ICQ Fussy-difficult	14.83 (5.00)	
ICQ Unpredictable	6.15 (2.66)	
ICQ Unadaptable	9.01 (3.83)	
ICQ Dull	5.84 (2.46)	
Strength and Difficulties Questionnaire		7.98 (4.63)
Child-Parent relationship (CPR- PIANTA), mean (SD)		
CPR- Positive		33.77
CPR- Conflict		15.64
Short Temperament Scale (LSAC),		
Persistence		4.71 (0.82
Sociability		4.12 (1.13)
Reactivity		2.88 (1.07)
Quality of attachment score (ASQ),	42.50 (2.59)	
Parental stress score, stressors	14.64 (4.19)	12.35
PCG Depression score, mean (SD)	2.49 (3.66)	2.42 (3.58)
Gender of PCG, n (%)		
Male	41 (0.4)	161 (1.6)
Female	11,093 (99.6)	9632
Global health Rating of PCG, n (%)		
Excellent/very good	7746 (69.6)	6760
Good/fair/poor	3387 (30.4)	3032
Ethnicity PCG, n (%)		
Irish	9275 (83.3)	8261(84.4)
White non-Irish	1203 (10.8)	1018
Black	295 (2.7)	252 (2.6)
Asian	273 (2.5)	202 (2.1)
Other	53 (0.5)	54 (0.6)
Family Social Class, n (%)		
Professional/Managerial	5340 (48.0)	4553
Other non-manual/Skilled manual	3643 (32.7)	3233
Semi-skilled/Unskilled	1148 (10.3)	1061

Unclassified	1002 (9.0)	947 (9.7)
Education PCG, n (%)		
Lower secondary or less	1955 (17.6)	1361
Upper secondary	2806 (25.2)	3192
Non-degree	3112 (28.0)	2080
Third level	3249 (29.2)	3144
Equivalised Annual Income, mean (SD)	21,507 (13,414)	17,874 (9,551)

PCG, primary caregiver.

*Hospital admission (ever): at 9 months this question related to an illness or health problem whereas at 3 years of age was for accident or injury.

Table 3.1 (continued) Demographics and profile of Growing Up in Ireland Infant Cohort Study Participants

3.3.2. Classification tree at 9 months of age

The results from CTA for infants at 9 months are shown in Figures 3.1 and Figure 3.2 which included the independent (predictor) variable used for each split. The tree model, without boosting or undersampling, had a sensitivity of 31.2%, a specificity of 90.4% and an overall accuracy of 88.8%. Boosting and resampling did not greatly improve the model performance for either dataset. The parameters used for growing the tree with the CHAID algorithm partitioned the data into 5 levels with 33 nodes of which 20 were terminal nodes. Each node contained the node number, the number and percentage of infants in each category for the dependant variable (dental problem), the adjusted p-value and chi-square statistic, the categories chosen by CHAID for the predictor variable and the cut-off points for continuous variables (Figure 3.1.2). The six independent variables that reached significance in the model included three infant temperament subscales (ICQ subscales *Fussy-difficult, Dull* and *Unpredictable*), the PCG depression score, infant use of a soother/dummy and child global health.

Four subscales are used for ICQ (*Fussy-difficult*, *Unadaptable, Dull* and *Unpredictable*) which are rated on a Likert scale with a higher score indicating an increased level of perceived difficulty in dealing with the behaviour described. The first level of the tree was split according to the infant temperament score (ICQ subscale *Unpredictable*) which split the tree root (parent node) into four branches (child nodes) with infant temperament scores

between 6.0-7.0 and greater than 8.0 ending in two terminal nodes (no further splits). Infants with an ICQ subscale *Unpredictable* score between 7.0-8.0 (node 3; 8.3% of total sample) had a dental problem prevalence of 5.7% and was split at the second level by whether or not they had used a soother/dummy. Those who used a soother/dummy had a dental problem prevalence of 7.0% compared to 3.0% for those who did not. Below this side of the tree those who had a habit of soother/dummy use were split by the *Dull* subscale of ICQ at level three (node 8; 5.6% of total sample) and, in general, where *Dull* subscale of ICQ was a predictor (levels 2, 3 and 4), those who had a higher *Dull* ICQ score had a lower prevalence of dental problems. For those who did not use a soother/dummy (node 7; 2.7% of total sample) the third level split was by infant global health with those who rated “Very good” having a lower prevalence of dental problems. It should be noted that the sample size for the minor class in these subgroups was very low.

Generally, throughout the tree, as the subscale score for ICQ *Unpredictable* or *Fussy-difficult* increased there was a relative increase in the proportion of infants with a reported dental problem while an increase in score for ICQ *Dull* tended to reduce the risk of having a dental problem. There was also a tendency towards a subtree replication problem with the ICQ temperament variables *Dull* and *Unpredictable*. On the other side of the tree the subset of infants with the lowest score for ICQ subscale *Unpredictable* (63% of total sample) also had the lowest prevalence of dental problems (2.3%) and was split at the second level by the PCG total depression score (node 1). Almost 10% of the total sample had a PCG depression score greater than 4.0, with 4.3% of the subset of infants in this group having a dental problem (node 6) whereas only 1.9% of those infants who had a PCG total depression score less than 4.0 (node 5; 53% of total sample) had a dental problem. Below this branch second and third level predictors were all ICQ temperament subscale scores. The largest subgroup terminal nodes were those split at level 3 (node 10; 40% of total sample) by ICQ *Unpredictable* with those scoring <4.0 having half of the prevalence of dental problems (1.2%) as those who scored >4.0. Again, at levels 4 and 5 it is important to cautiously interpret the significance of predictors as the sample numbers in the minor class are very low in some nodes.

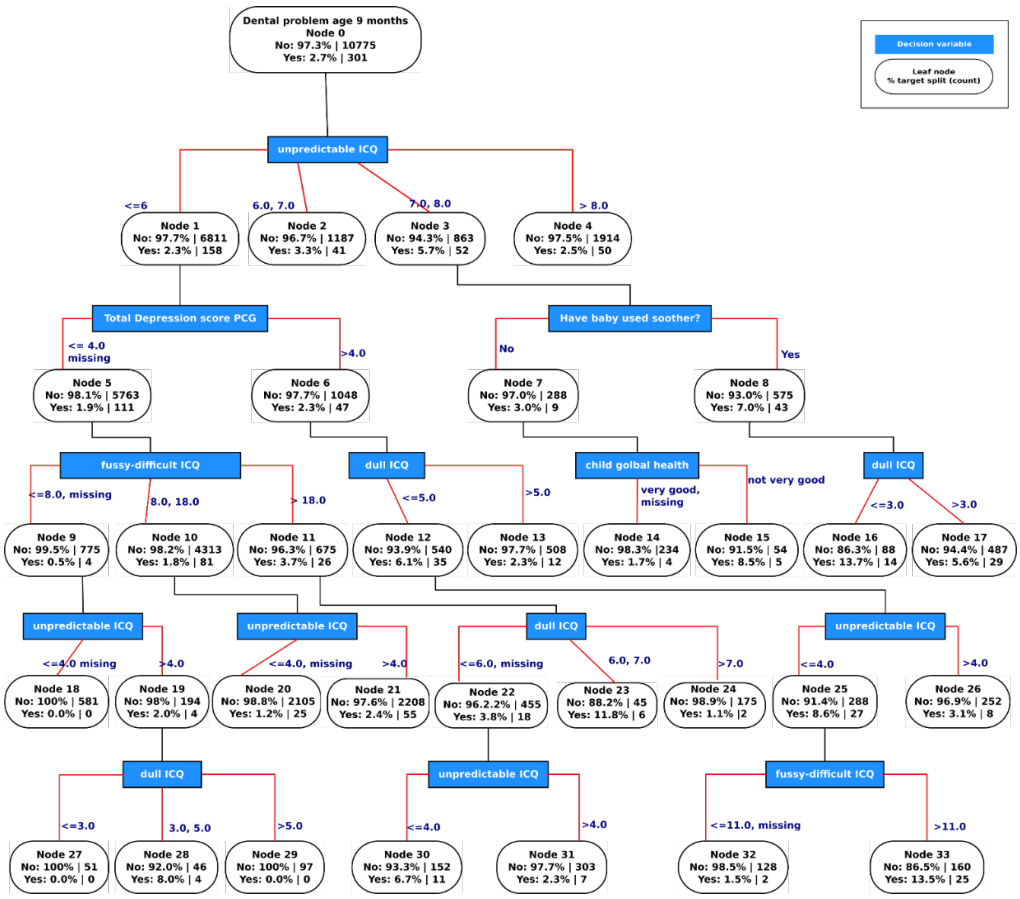


Figure 3.1 Prevalence of reported dental problems at 9 months of age among classification tree subgroups, percentage (%) and number (N) in each class.

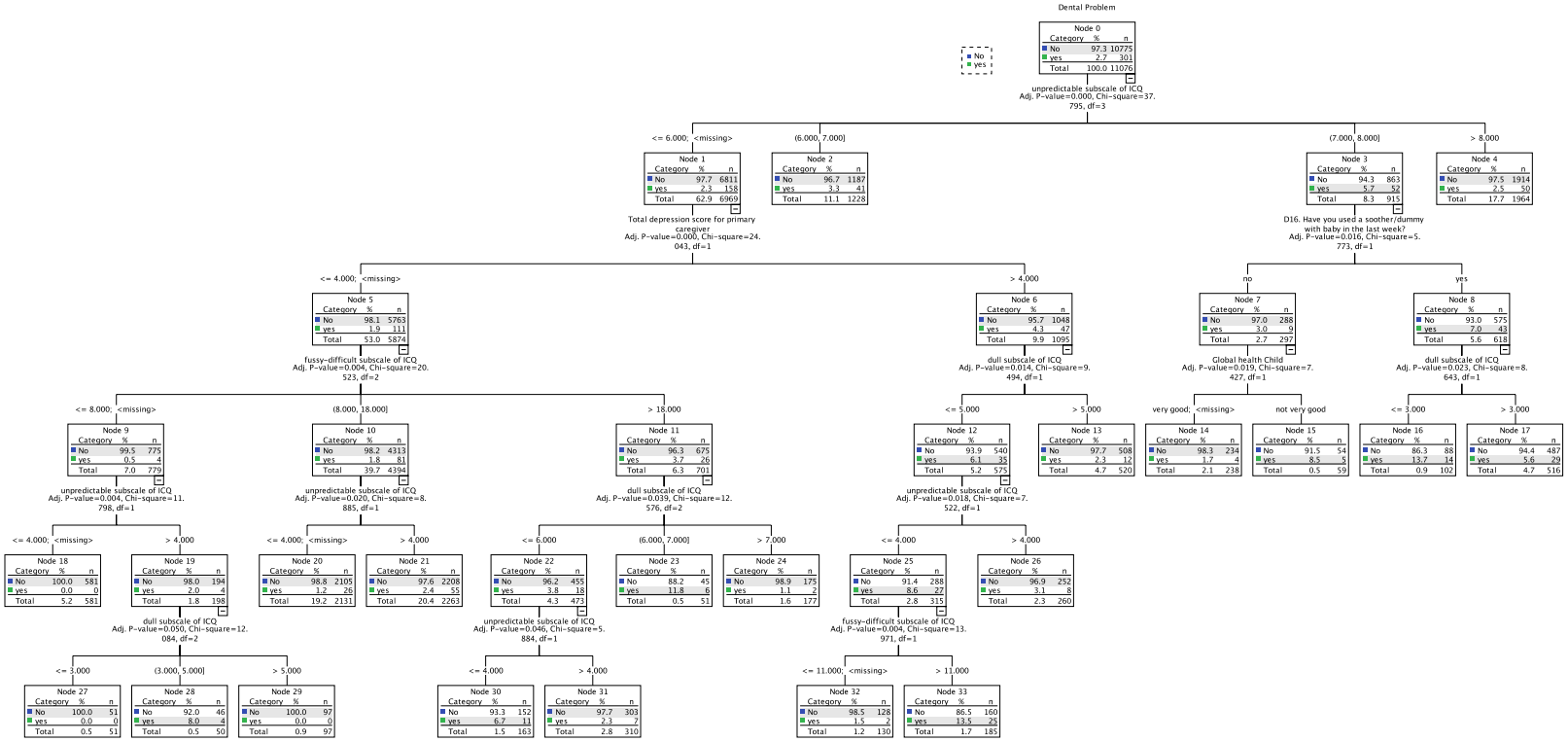


Figure 3.2 SPSS output showing prevalence of reported dental problems at 9 months of age among classification tree subgroups, percentage (%) and number (N) in each class, adjusted p-value and chi-square statistic.

In regression analyses three factors were significantly associated ($p<0.05$) with having a reported dental problem at 9 months: Both *Dull* and *Fussy-difficult* temperament (ICQ subscale scores), use of a soother and PCG depression score. Having a dull temperament or using a soother reduced the likelihood of a dental problem while having a difficult temperament and an increased PCG depression score increased the likelihood of the infant having a dental problem.

3.3.3. Classification tree at 3 years of age

The CTA for children at 3 years of age is shown in Figures 3.3 and Figure 3.4. This tree model resulted in a sensitivity of 66% and specificity of 58.5% and overall correctly classified almost 59% of all children in the dataset. This dataset was partitioned using the CHAID algorithm into 4 levels with 25 nodes of which 15 were terminal. Ten independent variables were significant in the model; child health (PCG global rating), ethnicity and education of the PCG, history of hospital admission for an injury, family annual income, PCG treated for depression, PCG stress score, total depression score, PCG health rating and *persistence* subscale of LSAC child temperament measure. The most important splitting variable was child global health rating and of those children who were classified as “Not very healthy” in total, 7.1% had a dental problem whereas of those who were classified as “Very healthy” only 4.3% had a dental problem. Approximately 75% (n=7274) of the total sample was classified as “Very healthy” and the next variable splitter at this node was PCG ethnicity. In the subset of children where the PCG was “other white background” (white non-Irish) the prevalence of dental problems was 8.4% whereas those of children of a PCG of Irish and “any other” ethnicity had a dental problem prevalence of 3.8%.

At the first level the node (2) representing the “Not very healthy” children split into two groups (Yes/No) on the basis of hospital admission for an accident or injury. Of those who had a hospital admission 11.1% had a dental problem compared to 6.2% for those who did not have a hospital admission. The second level splits in the tree were based on the highest education level of the PCG, whether PCG was treated for depression/anxiety, equivalised household annual income and PCG parental stress scores. At node 5 the prevalence of infant dental problems among families with an equivalised household annual income less than €11,800 was 4.0%, between €11,800 and €13,600 was 12.0% and greater than €13,600 was 5.7%. Only the lowest income group was split further according to the PCG global health rating.

On the “Very healthy” side of the tree those infants of PCG’s from “Any other white background” (non-Irish white) was split at the second level by PCG “treated for depression/anxiety” with infants of those who were treated having a dental prevalence of 16.4% compared to 7.6% for infants of those who were not treated for depression/anxiety. At node 3 the infants of those who were in the remainder of the ethnic groups were split into 3 subgroups by the PCG highest education level with no clear direction of association with dental problems. Only one of these groups had a final split at level 3 (node 8) according to PCG depression score with those greater than 4.0 having infants with a dental problem prevalence of 12.3% whereas those with a PCG depression score less than 4.0 had a dental problem prevalence of 6.2%.

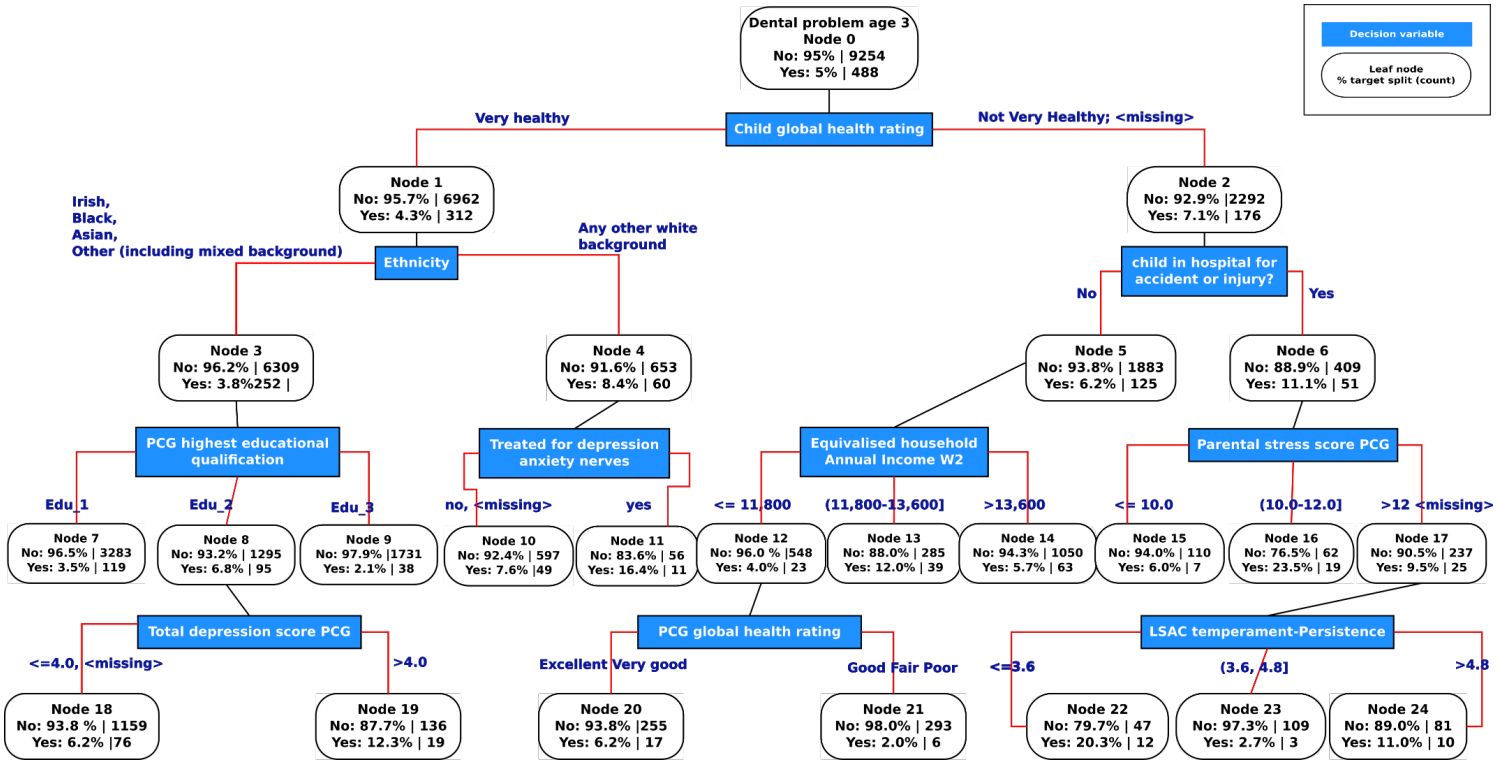


Figure 3.3 Prevalence of reported dental problems at 3 years of age among classification tree subgroups, percentage (%) and number (N) in each class.

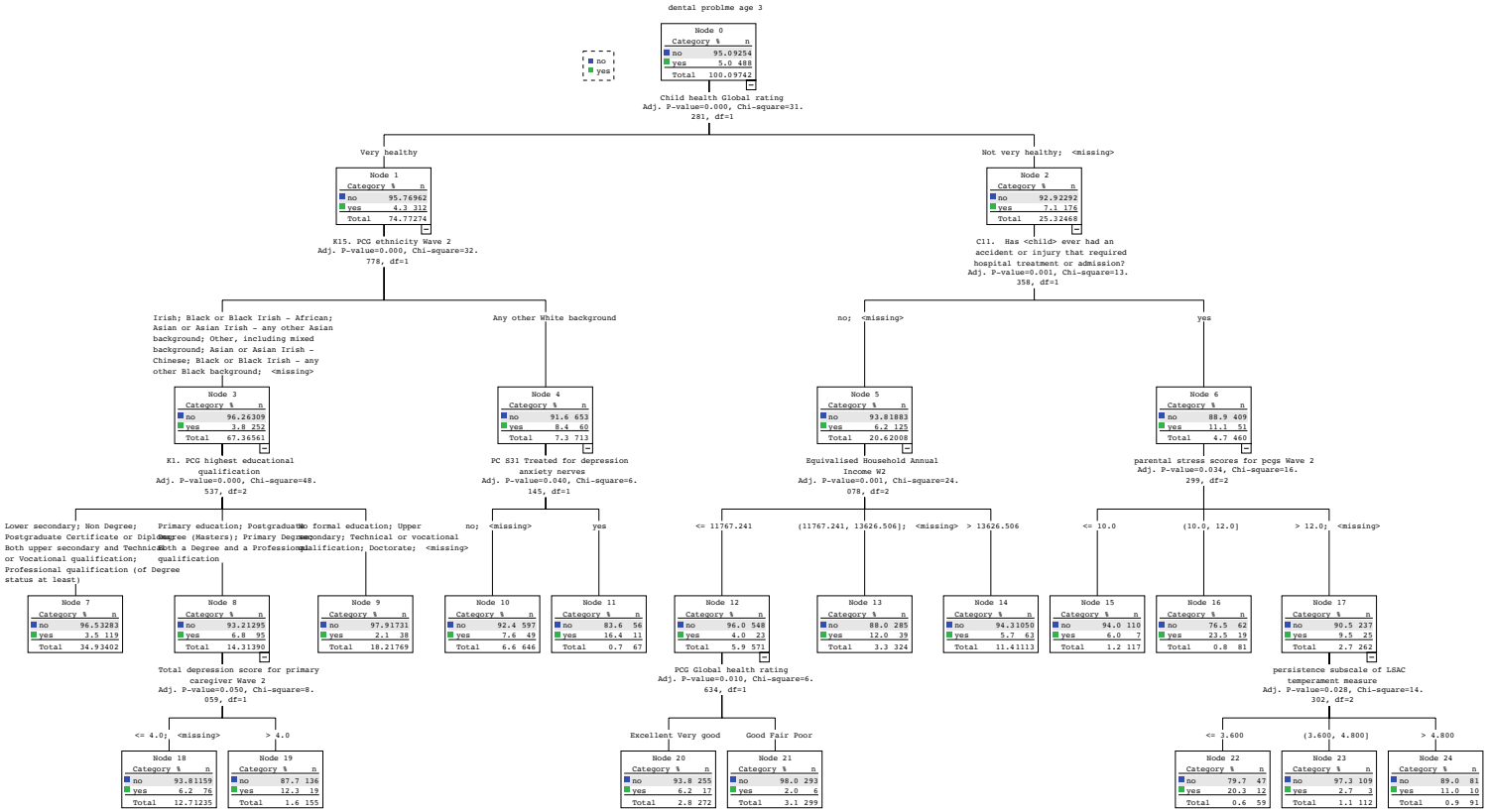


Figure 3.4 SPSS output showing prevalence of reported dental problems at 3 years of age among classification tree subgroups, percentage (%) and number (N) in each class, adjusted p-value and chi-square statistic.

In logistic regression analyses five factors were significantly associated ($p<0.05$) with the likelihood of being a child at 3 years of age with a reported dental problem: child global rating of health, hospital admission for injury, PCG ethnicity and PCG education level and family social class.

3.4. Discussion

The findings highlight the relatively high prevalence of reported dental problems in early childhood with one in fifty children at 9 months and one in twenty at 3 years of age with a reported dental problem which resulted in the PCG seeking child health care. Only 16 children had a dental problem at both time points. CTA showed that certain psychosocial characteristics, sociodemographic factors and measures of PCG and child health were key predictors of dental problems in preschool children. Specifically, infant temperament (*Unpredictable*) at 9 months and child health at 3 years were the most significant predictors of dental problems. While the logistic regression analyses largely supported these findings the CTA more clearly illustrated the complex multilevel interactions among a greater number of predictors.

This study provided exploratory analyses of the GUI infant cohort (at age 9 months and age 3 years) which is the largest child population study in Ireland covering key aspects of child health and development. Currently, national data is limited on child dental health and preschool children were not included in previous national dental surveys (Sagheri et al., 2013). To our knowledge no study has previously investigated the association between the prevalence of dental problems in preschoolers and health and psychosocial characteristics of the PCG and child using CTA. On a national population basis this represented approximately 2,000 infants age 9 months and 3,200 children age 3 years. This was similar to trends reported in other countries (Slack-Smith, 2003, Declerck et al., 2008). However, as our secondary analysis utilised a PCG reported dental problem as opposed to, for example, clinically diagnosed ECC, comparisons with other oral health indicators must be made with caution. Prevalence rates for ECC in 2 to 5 year olds from other European countries vary from approximately 6 to 28% (Baggio et al., 2015). The actual prevalence of dental problems in our cohort is possibly much higher given that only dental problems for which care was sought was recorded but not the actual outcomes of the visit. It is widely acknowledged that early childhood dental problems are

often neglected or not treated unless symptomatic (Sheiham, 2006, Slade, 2001).

The classification and assessment of important psycho-social variables is a common task in analysing data from large cohort studies related to child development. Typical statistical approaches include a univariate analysis and construction of a global model using a regression technique to see how well, or poorly, the parameters fit the model. The method of recursive partitioning in CTA facilitates visual identification of complex relationships in a large number of variables among subgroups of a population while not requiring the variable form and distribution assumptions necessary for parametric techniques (Lemon et al., 2003). CTA model multilevel interactions while regression methods largely assume that predictor variables act independently (Kuhn et al., 2014).

It is common when dealing with real problems such as medical or dental classification that the datasets are imbalanced (Rahman and Davis, 2013). Our study target variable involved a binary response with two classes where the one of interest was underrepresented which is described as the minority or positive class (Yap et al., 2014). Furthermore, the frequency of dental problems in the dataset for 9 month old infants was less than 5% which is described, statistically, as a rare event. Boosting and resampling techniques were applied to attempt to address the class imbalance but did not greatly improve the performance of either model. While class imbalance was greater in the 9 month dataset the main objective was to generate a classification tree for description of the characteristics affecting dental problems in the population subgroups. Overall accuracy of the model at 9 months was high (88.8%) but the sensitivity was low (31.2%) whereas the model at 3 years had high sensitivity (66.0%) and accuracy (58.9%). Given the problems of dealing with class imbalance the overall accuracy is often less useful an indicator of model performance than sensitivity (true positive rate) and specificity (true negative rate). A relatively high accuracy (88.8%) for the 9 month model belies the fact that the true positive rate (correctly predicting a dental problem) was only 31.2%. This has been referred to as the accuracy paradox, where the accuracy measure is high but only reflects the underlying class distribution. This emphasises the importance of reporting more performance indicators than accuracy alone. It would generally be more desirable to have a relatively high sensitivity and reasonably high accuracy to utilise the tree as a good predictive model (Hausen, 1997).

Social determinants and oral health problems are strongly related (Watt, 2007) as is supported by the results at 3 years of age. However, these factors were not significant predictors of infant dental problems at 9 months of age. Results from the Victorian Child Health and Wellbeing Study (Renzaho and Silva-Sanigorski, 2013) similarly found that socioeconomic factors were only significantly related to child oral health status in older children aged from 4-7 years. While the parental factors most commonly investigated that were associated with child oral health were sociodemographic there is an increasing awareness of the importance of the PCG and child psychological, general health and behavioural profiles as risk factors for ECC (Abreu et al., 2015). Our results suggest that the experience of dental problems at 9 months of age appears to be more related to infant temperament, use of a soother and PCG depression score.

The regression model also identified those infants at 9 months of age with a *Fussy-difficult* temperament as significantly associated with the prevalence of dental problems. Similar findings were previously reported with regard to increased prevalence of ECC and infants with a *Difficult* temperament (Spitz et al., 2006, Aminabadi et al., 2014). It may be that infant behaviours associated with a more *Difficult* temperament are interpreted as signs of hunger and those infants are fed more frequently. Poor infant feeding habits including a high frequency intake of 'sugary drinks' and 'night feeds' are high risk factors for ECC (Abreu et al., 2015); although waking infants at night for feeds was not a significant predictor in our classification.

Previous conceptual models have proposed that mothers with depressive symptoms and high parental stress may contribute to oral health problems in children through a number of pathways including early cessation of breastfeeding, unhealthy eating practices, poor oral hygiene behaviour and negatively affecting infant temperament (Kim Seow, 2012). The CTA (Figures 3.1 and 3.2) suggested that while PCG depression score was an important predictor (level 1) for dental problems at 9 months ($p<0.001$) for 63% of the sample population it was only a predictor at level 4 of the CTA when the infant was 3 years old ($p=0.05$). Results of regression analysis also found that PCG depression score was significantly associated with dental problems at 9 months but not at 3 years of age. However, whether the PCG was "treated for depression, anxiety or nerves" was a level 2 predictor ($p=0.04$) which split the children at 3 years who were from a non-Irish, white ethnic background (node 4).

It is important to note that as CHAID relies on contingency tables for calculating significance tests, continuous or ordinal variables must be coerced into a categorical form by binning. Thus, some continuous predictors were split by the algorithm at different cut points not necessarily related to clinical significance. For example, the CES-D depression score is a screening tool (increased score indicating increased likelihood of depressive symptoms) developed to measure depressive symptomology with "emphasis on the affective component, depressed mood" (Radloff, 1977) and is not a clinical diagnostic tool. Thus, CHAID does not determine the practical or applied cut-point for the CES-D depression score but automatically splits subgroup at that node according to the binning process. When the algorithm splits according to a predetermined categorical, binary, cut-point such as a 'depressed/not-depressed', then the actual score is not selected or coerced by binning.

The classification tree output at 3 years of age (Figures 3.3 and 3.4) showed that child global health was the most significant predictor of dental problem prevalence followed, at level 1, by PCG ethnicity and whether the child had a hospital admission for injury. Although child global health was a second level predictor at 9 months this was for a relatively small subgroup (node 7). Evidence for an association between a child's general and oral health continues to strengthen (Sheiham, 2006) and adverse childhood experiences, including psychosocial issues, are associated with poorer dental health (Bright et al., 2015). Previous national studies have used PCG-reported child health as a valid and 'more holistic' proxy measure for child health as defined by the WHO (Shrivastava et al., 2014). *The percentage of infants classified as "very healthy" at 9 months of age was almost 83% while at 3 years of age this had reduced to 75%. As both the prevalence of dental problems and of children classified as "not very healthy" increased at 3 years of age this may explain why child global health did not feature as a significant predictor at 9 months of age.* The current analysis strengthens the evidence for the association between overall child health and oral health in a nationally representative sample of preschool children.

In the current study, the "very healthy" children were split by PCG ethnicity and the infants of those from a non-Irish white background had a significantly higher prevalence of dental problems (8.4%) and this group was subdivided according to PCG being treated for depression or anxiety, which had almost double the rate of dental problems again (16.4%). It is not clear why this particular ethnic group appear to be more at risk for early childhood dental problems, but it may

be due to cultural differences related to caries risk factors such as diet, oral hygiene behaviours or PCG oral health beliefs (Kim Seow, 2012).

While PCG education level was the predictor that split the remainder of the "very healthy" subgroup it was difficult to interpret the results in a meaningful way. However, it was not surprising that PCG education and ethnicity were important components of the classification tree and regression analysis given the strong correlation between these sociodemographic factors and early childhood oral health (Abreu et al., 2015, Hooley et al., 2012b).

Results from the logistic regression analysis largely supported the results from the CTA in that, child health, PCG education, ethnicity, household income and family social class were significant factors ($p<0.05$) associated with reported dental problems at 3 years of age.

It was interesting to note that only 16 of the infants who had a dental problem at 9 months also had a dental problem at 3 years of age. While the nature of the dental problem may differ at these ages the associated predictors also varied with sociodemographic factors being more relevant at the older child age. This underlined the importance of a life course approach to investigating dental problems and using an exploratory analytical approach that can detect multilevel interactions (Ben-Shlomo and Kuh, 2002).

The GUI study is representative of the Irish population and a key strength is the range of detailed data collected that are useful in exploring predictors of dental problems in preschoolers. As far as we are aware, this is the first nationally representative study of this age cohort to investigate dental problems and psychosocial factors using CTA. The use of a graphical display tree allowed for easier visualisation of potentially important interactions and subgroups that might not be discovered using a more traditional statistical approach. Furthermore, as a non-parametric technique it is not restricted by the form and distribution of the variables being explored and doesn't require data transformations to utilise heavily skewed data.

It is important to acknowledge that although the classification tree method is useful from the clinician's perspective there are a number of limitations such as underfitting, overfitting and instability and the interpretation of the tree must be carried out with a degree of critical awareness of what may constitute a plausible relationship. The "oversensitivity" of CTA can cause small changes in input data to result in large changes to the tree appearance as all node splits

are dependent on the preceding splits (Kuhn et al., 2014). As indicated in Section 1.7.2 decision tree models can be cross validated, to minimise overfitting of the model, by dividing the sample into subsamples and trees generated with the data from each subsample excluded in turn (Ho Yu, 2010, Rokach and Maimon, 2009). A decision to manually stop splitting the tree or prune the tree by deleting nodes also helps limit the complexity of the tree and thus, reduce overfitting (Kingsford and Salzberg, 2008).

The data in the study is PCG-reported and consequently there is an increased risk of recall bias and social desirability. Although the psycho-social variables used are validated the study could be strengthened by including information relating to the children's dental condition. The outcome variable measured was PCG-reported dental problem requiring a health-care or dentist visit, which encompassed all oral health problems (including physiological problems such as teething), dental trauma and early childhood caries. Furthermore, the actual dental problems at 9 months of age can differ from those at 3 years of age when the child has a more developed primary dentition and risk factors for ECC are more likely to have an impact (Gussy et al., 2016, Wagner and Heinrich-Weltzien, 2017).

3.5. Conclusions

This study provides a clear visual representation, using classification tree analysis, of how PCG and child psychosocial and general health factors are associated with early childhood dental problems, even before the primary dentition is complete. The findings extend previous research by highlighting the relative importance of some known predictors and recognising the interconnected role of these factors in adopting an integrated multidisciplinary approach to assist in formulating a coherent oral health policy. CTA appears to be a useful, flexible and appropriate statistical approach to analysis of variable interactions in a large population-based research study without requiring particular distributional assumptions. Future research should focus on a life course approach to understand the multiple pathways through which these health and psychosocial factors in early childhood may impact on oral health throughout life.

Chapter 4. Weight status and dental problems in early childhood: classification tree analysis of a national cohort

4.1. Introduction

Early childhood caries is the most prevalent dental problem in preschoolers (Public Health England, 2013b), one of the most common causes of hospital admission and the most frequent reason for unplanned general anaesthesia in children (Gussy et al., 2006, Public Health England, 2013b). Obesity, defined as an excess of body fat (Flegal and Ogden, 2011), is another growing concern among preschool children. Body mass index (BMI) is frequently used to classify adults as overweight or obese; however, classifying overweight and obesity in children is complicated by age and gender specific differences (Flegal and Ogden, 2011, Rolland-Cachera, 2011). For this reason, the International Obesity Task Force (IOTF) defines childhood weight status based on BMI centile curves that correspond to adult criteria from 2-18 years for males and females (Cole et al., 2000). In Europe, 12-15% of preschool children are classified as overweight or obese based on IOTF criteria (Ahrens et al., 2014). Concerns around EEC and childhood obesity are heightened by the fact that both are strong predictors of these respective conditions throughout the life-course (Dye et al., 2015, Wake et al., 2008).

The preschool age is a particularly important period to minimise the risks for dental caries and obesity (Chaffee et al., 2015) and the primary caregiver (PCG) plays a key role in facilitating prevention through feeding patterns and other behaviours (Wake et al., 2008, Gussy et al., 2006, Dye et al., 2004). Obesity and dental caries share some common risk factors including food choice, dietary intake patterns, diet quality and socioeconomic factors such as PCG education and household income (Marshall et al., 2007b, Kantovitz et al., 2006, Sheiham and Watt, 2000). Given the associations that exist between oral health and general health interest is growing in using a common risk factor

approach to investigate the multidimensional causes of dental and weight-status problems, particularly in preschool children (Hooley et al., 2012a, Hooley et al., 2012c, Hayden et al., 2012, 2015a). Although some studies have shown a positive relationship between BMI and dental caries, others suggest that they are weakly correlated and that different predictors may be associated with dental caries at both high and low BMI levels (Hooley et al., 2012a, Kantovitz et al., 2006, 2015a, Marshall et al., 2007b). Indeed, very few studies report the oral health status of underweight children and often group underweight and normal weight without considering differences in risk (Tramini et al., 2009, Hooley et al., 2012a).

Data-driven methods are being increasingly proposed to empirically derive dietary patterns associated with chronic disease (Krebs-Smith et al., 2015). Methods that aim to uncover the relationship between independent variables and a dependent variable are described as supervised learning. The discovered relationship is typically presented as a classification or regression model (Yoo et al., 2012). Thus, in Classification and Regression Tree Analysis (CART) when the target (dependent) variable is continuous a regression analysis is performed and when the target variable is categorical a classification tree analysis (CTA) is carried out. Data mining techniques are invaluable when analysing multidimensional data from large-scale survey microdata files as they provide a means to identify novel diet-disease relationships and can help establish inter-relationships between causal factors (Yoo et al., 2012).

With few exceptions, most national dental surveys tend to focus on children aged 5 years and older. While nationally representative studies of obesity prevalence in older Irish children are well documented (Keane et al., 2014) there are few, apart from a National Preschool Nutrition Survey (Walton, 2012) that relate to preschoolers. The “Growing Up in Ireland” (GUI) study is a large, nationally representative cohort of 3-year old preschool children which does contain data related to weight status and dental problems. The research in this secondary analysis proposed to use a flexible analytical approach (CTA) to explore the multilevel relations between weight status and dental problems in this GUI cohort.

4.2. Methods

Data used for this analysis was derived from the GUI study when the children were 3-years old (Wave 2). BMI was calculated as weight divided by height squared (kg/m^2) and, for children, classified as overweight, obese, normal weight or thinness according to the IOTF age and gender specific cut-offs for 3-year olds (Cole et al., 2000, Cole et al., 2007). Overweight and obesity for children were also classified using the UK adaptation of percentile cut-offs from the WHO Multicentre Growth Reference Study (MGRS) with overweight criteria defined as a BMI between the 91^{st} and 98^{th} percentile while obesity was defined as a BMI on or greater than the 98^{th} percentile (Cole et al., 2012, Flegal and Cole, 2013). LMS parameters were created using Cole's LMS method (Cole, 1990) and PCG BMI was also categorised.

The dichotomous target variable was a PCG reported dental problem. Attributes (independent variables) that were relevant to the target variable (dependent variable) were selected for inclusion in the model based on findings from Chapter 3. Dietary intake (Marshall et al., 2007b, Chaffee et al., 2015, Harris et al., 2004) was assessed using a modified version of the Sallis-Amherst Food Frequency Questionnaire from the Longitudinal Study of Australian Children (LSAC) (Sallis et al., 2002).

Classification tree analysis was carried out using the following parameters in either SPSS (v. 20.0: SPSS, Chicago, IL) or SPSS Modeler (IBM SPSS Modeler v. 14.2: Chicago, IL) using the Chi-squared Automatic Interaction Detection (CHAID) algorithm (Kass, 1980): maximum tree depth=5, parent node=100, child node=50 and Bonferroni-adjusted chi-square statistic, significance <0.05.

4.3. Results

4.3.1. Cohort profile

Five percent of 3-year olds had a dental problem. As is common in investigations of health outcome the class distribution of the dataset was imbalanced. The minority class was the positive instances of having "a dental problem" and the negative response was the majority class.

Table 4.1 describes the cohort characteristics, including anthropometric measurements, child health and behaviours. Almost all of the self-identified PCGs were female and the biological parent of the study child. Eighty five percent were “Irish”. Using the IOTF cut-offs (Cole et al., 2000) the prevalence of thinness and obesity were 5.7% each with an additional ~18% of children being overweight. Using the WHO growth charts and BMI cut-offs, the prevalence of obesity was 12.8% with an additional 18.5% overweight.

Table 4.1 Weighted * Sample Characteristics, Growing Up in Ireland infant cohort participants 2010/11 (Child 3-years of age).

	Child		PCG	
			Mean	**SD**
Age (years)			29.6	(6.1)
Gender	**n**	**%**	**n**	**%**
Male	5024	51.3	161	1.6
Female	4769	48.7	9632	98.4
Anthropometrics	**Mean**	**SD**	**Mean**	**SD**
Weight (Kg)	15.27	(2.02)		
Height (m)	95.48	(3.92)		
Body Mass Index (Kg/m^2)				
Total	16.71	(1.61)	25.99	(5.16)
Male	16.99	(1.52)	26.98	(5.59)
Female	16.71	(1.61)	25.97	(5.15)
BMI Categories	**n**	**%**	**n**	**%**
Thinness IOTF	557	5.7	166	1.7
Normal IOTF	6685	68.3		
Normal WHO	6464	66.0	4523	46.2
Overweight IOTF	1737	17.7		
Overweight WHO	1815	18.5	2941	30.0
Obese IOTF	559	5.7		
Obese WHO	1257	12.8	1655	16.9
Missing	256	2.6	508	5.2
Child health and behaviours	**n**	**%**		
Dental Problems (in last 12 months)	493	5		
Longstanding illness or disability	1543	15.8		
Hospital admission (ever)	1569	16.1		
Tooth brushing 2 or more per day	5107	52.2		
Tooth brushing <2 per day	4685	47.8		
Thumb sucking	765	7.8		
Soother	3163	32.3		
TV viewing 1 hour or less per day	3569	36.4		
TV viewing 2 hours or less per day	3587	36.6		
TV viewing 2 hours or more per day	2630	26.9		
Socio-demographics			**n**	**%**
Ethnicity				
Irish			8261	84.4
Non-Irish white			1018	10.4
Black			252	2.6
Asian			202	2.1
Other			54	0.6
Family Social Class				

Professional/Managerial	4553	46.5
Other non-manual/Skilled manual	3233	33.0
Semi-skilled/Unskilled	1061	10.8
Unclassified	947	9.7
Highest Education level		
Lower secondary or less	1361	13.9
Upper secondary	3192	32.6
Non-degree	2080	21.2
Third level	3144	32.1
	Mean	**SD**
Equivalised Annual Income (€)	18,004	(10,997)

Data presented as mean and standard deviation (SD) or n and percentage. *Sample weighting factors applied to statistically adjust the data to be more representative of the population. IOTF, International Obesity Task Force; WHO, World Health Organisation.

Table 4.1 (continued) Weighted * Sample Characteristics, Growing Up in Ireland infant cohort participants 2010/11 (Child 3-years of age).

4.3.1. Dietary Intake

The frequency of food items consumed is reported in Figure 4.1. The majority of children consumed water (~83.0%), full-fat milk/cream (~84.5%), full-fat cheese/yoghurt (~85.0%), cooked vegetable (~85.0%), fresh fruit (~89%), and biscuits/doughnuts/cake/chocolate (~74%) once or more than once in the previous 24-hours. Of interest, a considerable proportion of 3-year olds consumed “un-healthy” foods including crisps (~47%), hot-chips (~28%), sugar containing drinks (~30%), and sweets (~49%).

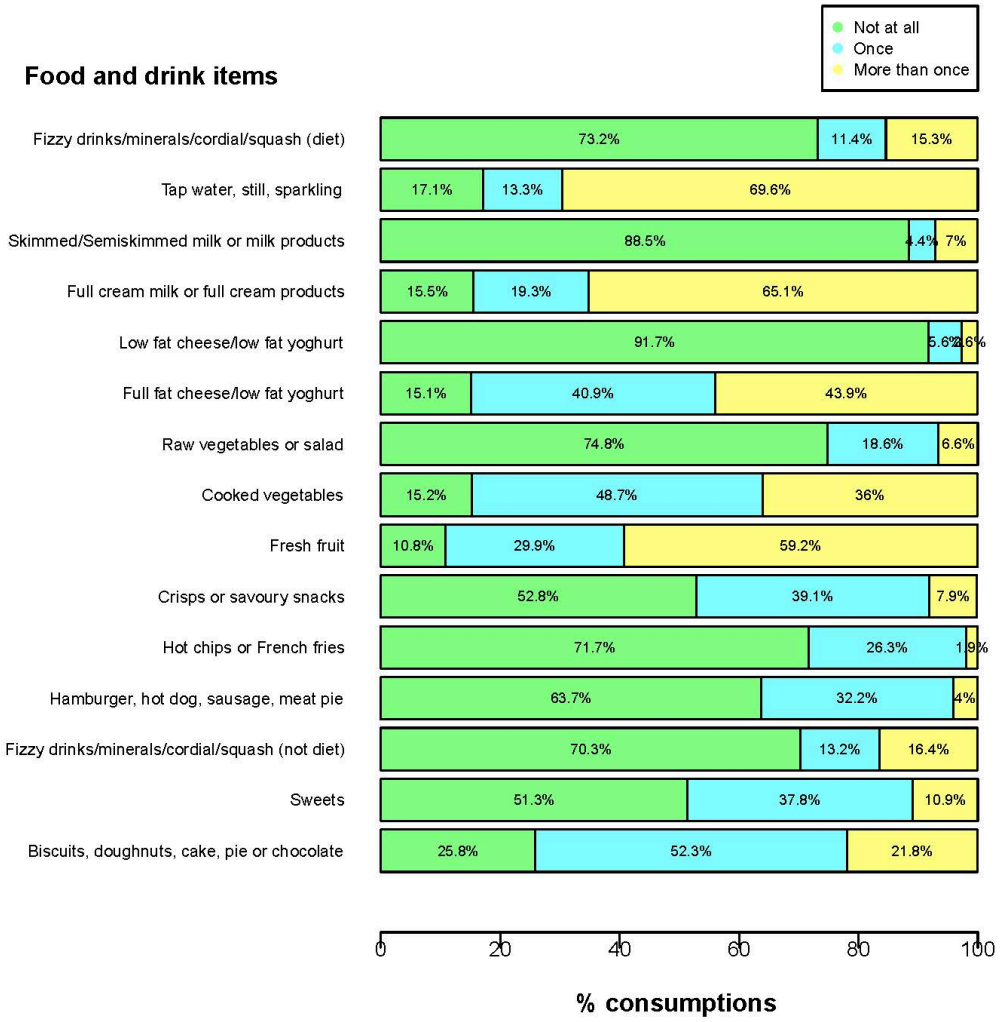


Figure 4.1 Food and drink items consumed in the previous 24 hours by the Growing Up in Ireland infant cohort at 3-years of age. For example, almost 38% of children consumed sweets once while almost 11% consumed sweets more than once in the previous 24 hours.

4.3.2. Classification Tree Analysis

CHAID analysis generated a CTA output as depicted in Figure 4.2 with 30 nodes, including 17 terminal nodes. Each node contains the number and percentage of infants in each category for the dependant variable (dental problem), the categories chosen by CHAID for the predictor variable and the cut-off points for continuous variables. PCG ethnicity was the most important predictor of the 3 year old child having a dental problem splitting the root node. Twelve predictor variables were included in the final tree (Bonferroni-adjusted

p<0.05). Two predictors appeared twice in the output, PCG BMI (nodes 2 and 5) and equivalised household annual income (nodes 3 and 4). A confusion matrix (Table 2) produced performance metrics for the classification tree: sensitivity 66.8%, specificity 58.5% and overall accuracy 58.9%.

The ethnic subgroups were split into 3 nodes with the highest prevalence of dental problems (8.4%) among those children from a “non-Irish white” background (Node 3). Node 1 contained almost 87% of the sample (Irish and Asian ethnicity) with a 4.7% prevalence of dental problems. The tree output from node 1 to nodes 22-24 delineated subgroups linking child BMI categories with dental problems by the following predictors: PCG from an Irish/Asian background (node1), the presence of a longstanding illness or disability in the child (node 5) and an overweight mother (node 13). The final predictor at node 13 was BMI classification of the child, which split into three terminal nodes resulting in normal, overweight/missing and obese/underweight subgroups. The highest dental problem prevalence (19%, n=17) was in those children in this final subgroup who were obese or underweight (node 24). In addition, the subgroup at node-1 who had a longstanding illness or disability had a reported dental problem prevalence of 7.0% while those with no illness or disability had a prevalence of 4.3% (node 4). The food variables included in the tree output were water (level 3), low-fat cheese/yoghurt (level 4) and raw vegetables/salad, fresh fruit and hot chips. Logistic regression failed to generate a significant model (chi-square (6)=9.38, p=0.15).

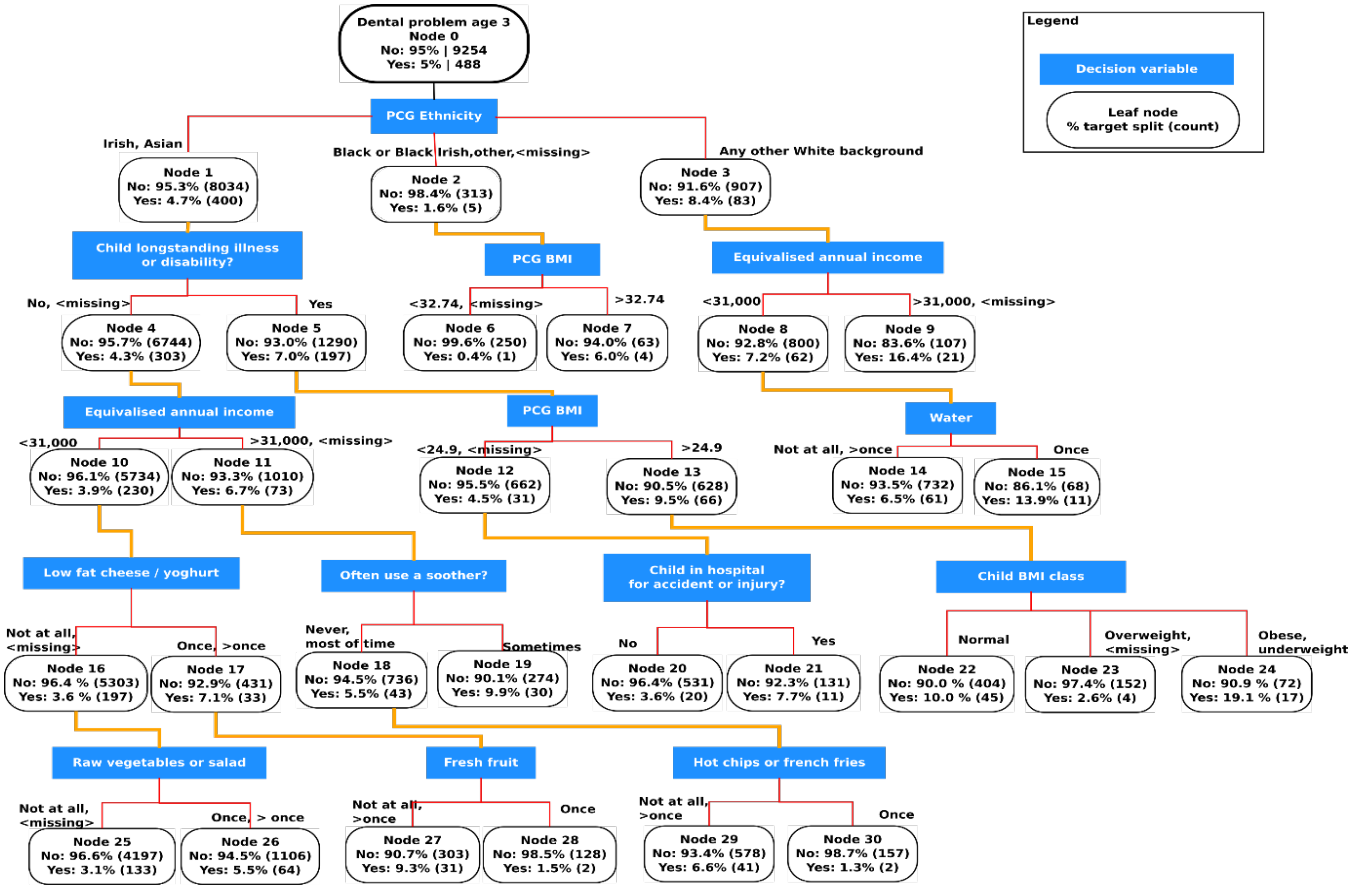


Figure 4.2 Prevalence of reported dental problems by the Growing Up in Ireland infant cohort at 3-years of age among classification tree subgroups, percentage (%) and number (n) in each class.

Table 4.2 Confusion matrix showing selected performance measures for Classification tree analysis of dental problem prevalence in the Growing Up in Ireland infant cohort at 3-years of age.

Observed		Predicted Dental Problem		Percentage Correct	Measure
		Yes	No		
Dental problem	Yes	326	162	66.8%	Sensitivity [a]
	No	3839	5415	58.5%	Specificity [b]
Overall percentage				58.9%	Accuracy [c]

[a] Sensitivity = True Positive Rate = Number of True Positives/(Number of True Positives + Number of False Negatives); [b] Specificity = True Negative Rate = Number of True Negatives/(Number of True Negatives + Number of False Positives); [c] Accuracy = (True Positives + True Negatives)/(True Positives + False Positives + False Negatives + True Negatives).

4.4. Discussion

This study used a CHAID classification tree as a method to classify the dataset and identify relationships between the predictor variables selected and the target variable (a PCG-reported dental problem requiring a visit to the dentist). PCG ethnicity was the most significant predictor of dental problems in the CTA model and the highest prevalence of dental problems in this study was among children who were obese or underweight with a longstanding illness and an overweight PCG.

This analysis was carried out using data from a nationally representative cohort of 3 year old children, the largest child population study ever carried out in Ireland, which includes a wide range of PCG and child health and development characteristics. CTA is a non-parametric method which handles nominal and numeric input and classification trees are ideal for representing complex interactions (Yoo et al., 2012). The output produces a visualisation of all the significant interactions with the target variable at multiple levels and can, potentially, uncover subgroups that might not be discovered using other data

methods. Classification and decision trees can also handle data with missing data and errors (Maimon and Rokach, 2005).

However, there are limitations in interpreting main effects (Maimon and Rokach, 2005). Given the hierarchical nature of recursive partitioning CTA has an inherent instability to small changes in the learning data. Overfitting of the model is a known problem in CTA and this can be guarded against by using cross-validation which splits the dataset into a training portion and test portion before estimating an average misclassification risk (Yoo et al., 2012). The problem of handling imbalanced class distributions when investigating health outcomes is discussed in the next section. While dental caries is the most common dental problem at this age the survey did not include a report of the outcome of the dental visit. However, it has been shown that dental disease and treatment need of young children are associated with parents' perceptions of their children's oral health status (Talekar et al., 2005). This study was based on a cross sectional rather than longitudinal analysis of the infant cohort from the GUI survey. The data is reliant on PCG reporting and this is subject to recall bias and social desirability, particularly in relation to reporting of food and drink perceived as 'healthy' or 'unhealthy'.

It is important to understand both the type of data mining algorithms available before applying them and evaluate the quality of the input data selected to avoid the risk of adopting a "black-box approach" (Breiman, 2001). This analysis used CHAID which uses Bonferroni-adjusted significance testing and can be used for prediction or classification and interaction between variables (Yoo et al., 2012). While the results of our CTA were exploratory they identified certain characteristics previously suggested as risk factors or risk indicators for both obesity and dental caries (Harris et al., 2004, de Onis and Lobstein, 2010, Hooley et al., 2012a). The study supports recent views that data driven-outcome dependant methods such as CTA are potentially useful for investigating dietary components or patterns most associated with a health outcome and are a valid, non-parametric, alternative to logistic regression analysis (Krebs-Smith et al., 2015).The results of this analysis must be cautiously interpreted by gauging the model performance (Table 4.2) and understanding the limitations of both the data structure and classification tree algorithms. An imbalanced class distribution has been characterised as "one that has many more instances of some classes than others" (Sun et al., 2007). CTA of imbalanced data sets tends to result in high predictive accuracy for the majority class and low accuracy for the minority class. The algorithms generally

favour high overall classification without any regard for individual class significance (Maimon and Rokach, 2005). In most health outcome investigations, including dental problems, the correct classification of the minority class is of greater interest or value than that of the majority class. The confusion matrix (Table 2) shows the results of the actual and predicted classifications carried out by CTA. The metrics calculated include sensitivity or recall (66.8%) which is the proportion of actual positive cases correctly predicted by the model and specificity (58.5%) which is the proportion of actual negative cases correctly identified by the model. The overall accuracy (58.9%) indicated the proportion of the total number of correct predictions. Logistic regression did not perform well as a classifier and none of the same input variables were significant in the final regression model. To be suitable as a prediction model for targeting risk it has been suggested that both sensitivity and specificity should be 80% or the sum be at least 160% (Hausen, 1997). While the CTA model performed well compared to logistic regression as a classifier the metrics achieved would not be sufficient to consider it appropriate for prediction or prevention at the level of the individual.

The partitioning variable at the first level of the classification tree was PCG ethnicity while at the next level the most significant predictors of dental problems were the child having a longstanding illness, PCG BMI and household income. It has been reported that trends in overweight and obesity differ among different ethnic groups, even at the early preschool age, and that this cannot be explained by variations in household income (Karlsen et al., 2014). Similarly, the disparities in dental caries prevalence between different ethnic groups is not fully explained by social inequalities (van der Tas et al., 2016). Surprisingly, ethnicity has been used as an independent variable in relatively few studies of dental caries in children while the PCG education level, although it was not a significant predictor in our CTA model, has been consistently shown to be an important risk factor for caries in children (Harris et al., 2004).

While overweight and obesity dominate the focus of recent research with children, it is also important to consider underweight (thinness) in early childhood as a condition related to poor health outcomes. There is some evidence to suggest that dental caries may be associated with children who are underweight and suffer with slow growth due to pain on mastication (Hooley et al., 2012c, Sheiham, 2006, Tramini et al., 2009). The results (Table 4.1) showed a similar prevalence of underweight and obese children. A small subgroup (node 24) of children which combined obese and underweight

categories had the highest prevalence of dental problems (19%) in the sample. This group were predominantly Irish with a longstanding illness and had an overweight PCG. It should also be noted that normal weight children (node 22) in this group had a dental problem prevalence of 10%; approximately half that of the obese/underweight group, but double that of the overall sample. This finding of itself highlights the interconnected nature of weight status, dental problems and general health and reinforces the importance of adopting a common risk factor approach when dealing with prevention of these diseases (Sheiham and Watt, 2000, Peres et al., 2016). However, while of interest in classifying this dataset, it is important to be cautious when interpreting these subgroups identified by CTA as hierarchical splitting means that they are mutually exclusive. Furthermore, successful targeting of high risk population subgroups for problems with both weight status and dental health would require a risk prediction model with both high sensitivity and specificity.

The prevalence of PCG-reported dental problems requiring a visit to the dentist was 5% which may be an under-estimation given that dental problems are often not treated in the preschool years unless symptomatic (Sheiham, 2006). This age is a pivotal period for development of both obesity and dental caries as patterns of eating behaviour that predispose to later development of these conditions are established (Wake et al., 2008, Dye et al., 2004, Hooley et al., 2012c). The prevalence of overweight or obesity in 3 year old children determined by IOTF cut-offs was approximately 23% which was similar to previous reports (Walton, 2012). Almost 47% of PCGs were overweight or obese and it is well established that parental overweight and obesity increases the risk of a child becoming overweight (Lobstein et al., 2015). There are limitations in using BMI as an indirect measure of “fatness” particularly with respect to children (Rolland-Cachera, 2011) and it is important to note that there is no reference population in Ireland for grading BMI. In the CTA, the IOTF classification was used as the more conservative estimate of obese children with a higher cut-off threshold. The FFQ adopted for the GUI survey was a modified dietary screening recall and provided an indication of types and frequency of foods consumed. While PCG-reported measures of foods consumed on a single occasion may be useful in differentiating patterns of food intake it does not provide a good estimate of usual daily consumption and cannot accurately capture total energy or total nutrient intake (Magarey et al., 2011). Preschool children with unhealthy eating habits have an increased likelihood of experiencing dental caries (Dye et al., 2004). While obesity and

dental caries are both diet-mediated diseases it is clear that sugars are required in the diet for dental caries to occur (Peres et al., 2016) whereas a high consumption of energy dense foods including sugars and saturated fats are linked with obesity (Hayden et al., 2012, Lobstein et al., 2015). Fundamentally, obesity occurs due to an energy imbalance between calories consumed versus those expended over a period of time (Marshall et al., 2007b, Lobstein et al., 2015). Approximately 74% of the children in GUI consumed biscuits, doughnuts, cake, pie or chocolate at least once or more than once in a 24- hour period (Figure 4.1). Almost half of the children ate sweets and 30% drank non-diet fizzy drinks, minerals, cordials or squash at least once or more than once.

Further research should be focussed on maximising the quality of food intake data by augmenting the GUI survey data with more reliable dietary intake values from a national nutritional database (See Section 5). Inclusion of more detailed oral health measures should also be considered.

4.5. Conclusions

The highest prevalence of dental problems in this study was among children who were obese or underweight with a longstanding illness and an overweight PCG. Societal changes may require renewed focus on oral health policies to focus on minority groups and CTA is a novel approach for exploring large survey data and health-related outcomes. The common risk factor approach may be a pragmatic means of developing shared modifiable strategies for prevention of both dental and weight problems.

Chapter 5. Data mapping protocols to augment the quality of reported food intake data in a short food questionnaire

5.1. Introduction

Exploring potential diet-disease relationships requires an accurate estimate of food intake. The difficulties associated with measuring diet are well documented (Foster and Adamson, 2014, Biro et al., 2002, Faber et al., 2013, Thompson and Subar, 2013, Satija et al., 2015). Collecting accurate and detailed dietary intake data is costly at a national level, therefore, dietary assessment tools are often modified or limited accordingly (Faber et al., 2013, Golley et al., 2017). While all dietary assessment methods are prone to measurement error (Carroll et al., 2012, Thompson and Subar, 2013) there are a number of factors to consider when selecting the most appropriate method, particularly for young children where the primary caregiver (PCG) usually provides a proxy report of food intake (Magarey et al., 2011). Firstly, it is important to consider which aspects of the diet are of interest such as specific foods, episodically consumed foods, or total food and nutrient intake, while study design and objectives will also influence the method selected (Table 1.4). In large-scale cohort surveys dietary intake is often assessed to either describe usual intake distributions or estimate the relationship with a particular health outcome.

Food Frequency Questionnaires (FFQ), 24 hour recalls, multiple-day food diaries (FD) or records, diet histories, and biomarkers are some of the most commonly used methods to assess dietary intake (Golley et al., 2017, Shim et al., 2014, Biro et al., 2002, Rutishauser, 2007, Thompson and Subar, 2013). Smaller studies tend to use prospective methods such as the detailed weighed FD over a number of days or weeks which can estimate the distribution of

habitual intake of a group (Rutishauser, 2007). Despite having limitations, the weighed FD method is considered the “standard” reference for relative validation in nutritional research (Satija et al., 2015, Taren et al., 2006). In addition to the self-report methods described, there are a number of dietary biomarkers that reflect nutrient and food intakes; for example, serum vitamins, blood lipids and urinary electrolytes (Jenab et al., 2009, O'Sullivan et al., 2011). Comprehensive reviews of the different methods and their limitations and strengths have been widely reported (Foster and Adamson, 2014, Biro et al., 2002, Faber et al., 2013, Thompson and Subar, 2013, Satija et al., 2015, Shim et al., 2014, Moynihan et al., 2009).

A FFQ is the most widely used dietary method for epidemiological studies and this is sometimes further modified in terms of time-frame, food items and estimation of quantity (Golley et al., 2017). However, the data generated is limited, particularly if key foods are omitted and minimal consumption frequencies recorded. For example, even relatively simple descriptive analysis of “unhealthy” food intake data can be compromised and bias our understanding of the potential association with chronic disease (Kirkpatrick et al., 2014, Thompson and Subar, 2013). Furthermore, measuring habitual food intake has a number of inherent issues such as self-selection, social desirability bias and selective underreporting of specific foods (Lissner, 2006, Magarey et al., 2011, Thompson and Subar, 2013). Short Food Questionnaire (SFQs) are increasingly used in national cohort surveys to measure aspects of dietary intake, however, publications rarely report details of relative validation or measurement error (Golley et al., 2017).

An “unhealthy” diet is a major factor that contributes to obesity, diabetes, cardiovascular disease and poor oral health (Anderson et al., 2016, Lobstein and Davies, 2009). Sugar intake is the most important risk factor for dental caries. Sugar containing foods and drinks are also targeted as a means to reduce total energy intake and therefore help control body weight and obesity (Te Morenga et al., 2013, World Health Organization, 2015, Moynihan and Petersen, 2004, Moynihan, 2016). Socioeconomic status is another common risk factor for dental caries and obesity (Public Health England, 2015a). Therefore, it makes sense to adopt a common risk factor approach to address both conditions, given the limited public health resources available (Chi et al., 2017, Public Health England, 2015a, Sheiham and Watt, 2000). Although there is a lack of good quality dietary data for preschool aged children (Amezdroz et al., 2015, Chankanka et al., 2015) there is evidence that total and free sugar

intakes in most developed countries exceeds the population dietary guidelines and the WHO recommended thresholds (World Health Organization, 2015). Free sugars include monosaccharides and disaccharides added to foods and beverages by the manufacturer, cook or consumer, as well as sugars naturally present in honey, syrups, fruit juices and fruit juice concentrates (World Health Organization, 2015). Recent studies have also indicated that this early preschool period may be a critical opportunity for early intervention to promote healthy growth, body composition and dental health. In particular, studies show that early changes in dietary behaviour can result in remineralisation of non-cavitated early lesions in teeth (Johnson et al., 2016, Gussy et al., 2016). Similarly, although there is a paucity of studies at this age, multicomponent programs to prevent or treat childhood obesity, particularly with parental involvement, have successfully impacted on pre-school child weight (Bluford et al., 2007).

GUI is a nationally representative longitudinal study of infants in the Republic of Ireland which used a SFQ (with no portion sizes) to assess the intake of "healthy" and "unhealthy" food and drink by 3-year old preschool children (Sallis et al., 2002, Quail et al., 2011). The NPNS provides the most accurate estimates available for dietary intake of young children in Ireland (Irish Universities Nutrition Alliance, 2012) using a detailed 4-day weighed FD. This chapter describes a method that can be used to link matched datasets to improve the quality of dietary data collected using SFQ's in large cohort surveys. It illustrates the application of this method using two national surveys: i) GUI, which collected food consumption data using SFQs and ii) NPNS, which collected food consumption data using a weighed FD. We report foods that were *covered* or *non-covered* by the SFQ in GUI relative to the detailed dietary assessment in NPNS. In this study the NPNS food database was used as the "reference standard" to map onto the larger national cohort survey and create an augmented food intake database. This study adds to previous reporting of the risk involved when selecting brief SFQs for large studies which may be less costly and less burdensome than detailed methods but increase the risk of attenuating the relationship between dietary factors and health outcomes (Faber et al., 2013, Golley et al., 2017, Shim et al., 2014, Taren et al., 2006).

5.2. Methods

This research utilised data collected as part of two nationally representative studies: the second wave of the GUI infant cohort longitudinal survey and the NPNS cross-sectional study. The second wave of the GUI infant cohort were 3 years of age at the time of interview (n= 9,793). The NPNS had a total sample of 500 children aged 2-4 years; but only the 3 year olds were included for this analysis (n=126).

In the GUI study, dietary intake was assessed using a SFQ, previously used in the Longitudinal Study of Australian Children (LSAC), to characterise healthy and unhealthy food intake (Sallis et al., 2002). No information on food portion size was recorded. A 4-day weighed food record was used in NPNS to collect food and beverage intake data (Irish Universities Nutrition Alliance, 2012). In total, there were 1,652 different food codes in the NPNS and each food was also assigned to one of 77 food group categories.

Data files were imported from SPSS (v. 20.0: SPSS, Chicago, IL) or converted to .csv format before importing to R (version 3.2.2) for linkage and analysis (Appendix). The NPNS Food group categories (n=77) were used for this analysis and other variables such as food name, cooking method, day of consumption, meal-type and food description were also selected. A unidirectional mapping procedure (Figure 2.3, Section 2.6.3) was carried out using a shallow natural language processing (NLP) approach.

All food categories in NPNS were sorted, grouped and filtered to facilitate easy mapping whereby all GUI food groups were filled with information from the NPNS food datafile and consolidated into a single augmented database. The augmented data were analysed to examine all food groups described in NPNS and GUI and what proportion of foods were *covered*, *non-covered* or *partially-covered* by GUI food groups relative to the NPNS database which included a more detailed dietary record. The term *non-covered* indicated a specific food consumption that could not be mapped using a GUI food code, i.e., the food in NPNS is not matched by the same food in GUI. Examples of both the food intake entries for a subject on a survey day (Table 2.2) and the manual mapping process illustrating *partially-covered* GUI food categories (Figure 2.4) are shown in Section 2.6.3.

The initial aggregation was done at the subject and survey day levels meaning that for each subject and each day of the survey an aggregated record was

obtained. The analysis treated each day of the 4 days in NPNS as an independent day. The mean daily intake amount (g/day) and the frequency of each food consumed was calculated for each NPNS participant, by summing the amount of all food consumptions a subject consumed per food group, averaging across the four days for each subject and then calculating the total sample average. Frequency was estimated by summing the total number of times the food appeared in the diary and dividing by four, i.e. the number of days in the survey.

Estimates were also derived for the percentage of consumptions per subject per day for each NPNS food group that was *non-covered* as a percentage of the total number of consumptions. A similar ratio was calculated for the percentage amount of food items *non-covered* over the total amount of food consumed per day. The total number of times when a *non-covered* food was consumed (total consumption frequency per day) and the total food amount (g/d) of a *non-covered* food was calculated. The ratio of the frequency of consumption of *non-covered* food over the total food frequency was determined. A similar ratio was determined for the amount of *non-covered* food consumed over amount of total food consumed. The frequency distributions of the ratio of consumption frequency and amount of *non-covered* food consumed divided by the total food consumed were displayed as histograms. Using a non-parametric density estimation, the distribution of the proportion of *non-covered* food consumed each day of the week was displayed graphically using kernel density plots and tested formally using the Wilcoxon rank sum test ($p<0.01$) (Dalgaard, 2008).

Further details are included in Section 2.6.5. and Appendix A.

5.3. Results

The unidirectional mapping protocol (Figure 2.3, Section 2.6.3) created an augmented food database which was then aggregated to produce quantitative metrics to assess how well the SFQ in GUI performed in matching a detailed national food database for the same age cohort in NPNS.

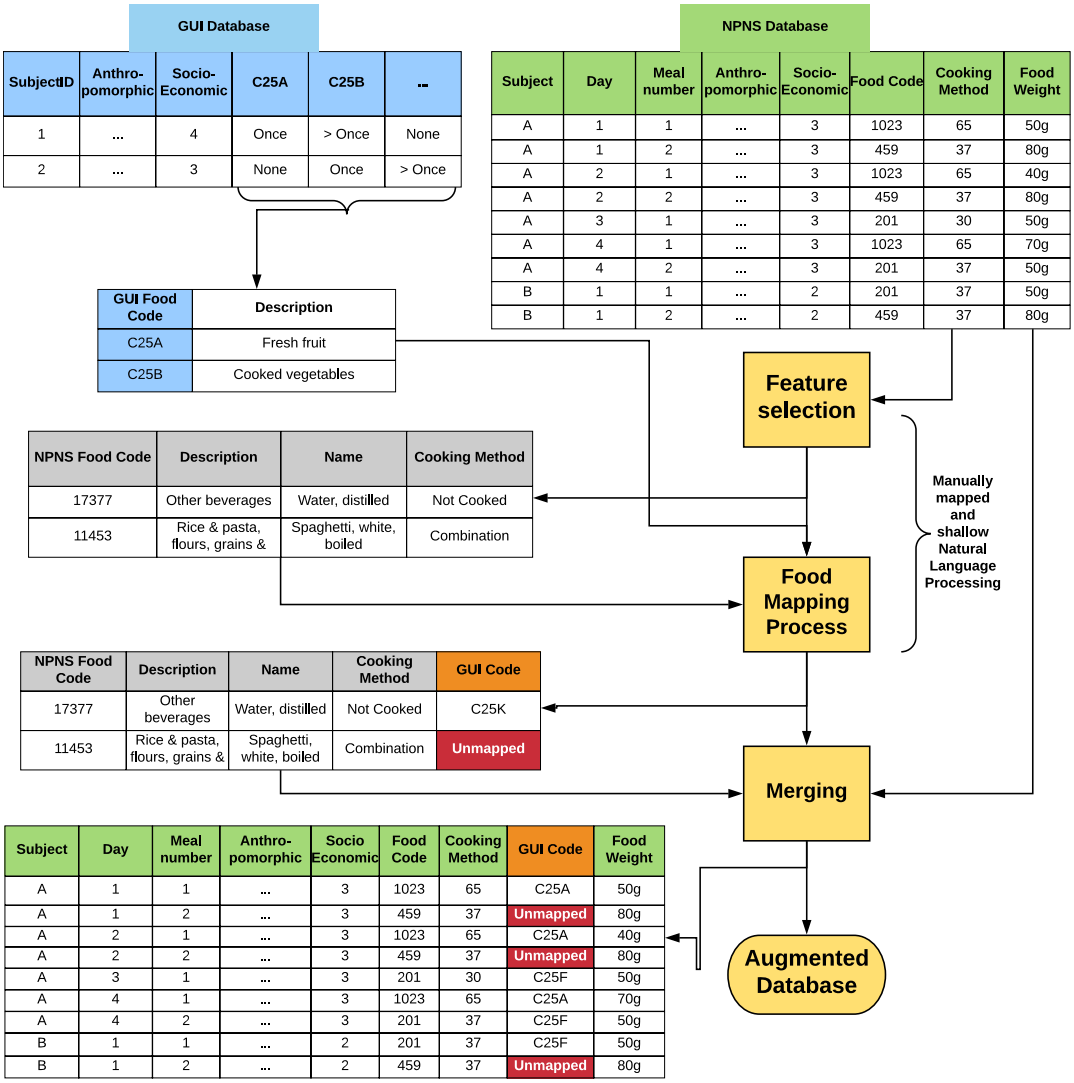


Figure 2.3 Flow diagram showing data processing steps for unidirectional mapping of GUI food codes with NPNS food codes. GUI: Growing Up in Ireland; NPNS: National Preschool Nutrition Survey. Feature selection identified variables from both GUI and NPNS databases that were desired, e.g. socioeconomic class, cooking method, food weight. All GUI codes were manually mapped with food categories from NPNS, e.g. NPNS food code 17377 mapped to GUI code C25k; NPNS food code 11453 was unmapped and this created a *non-covered* food group.

Characteristics of both the NPNS and GUI surveys are presented in Table 5.1.

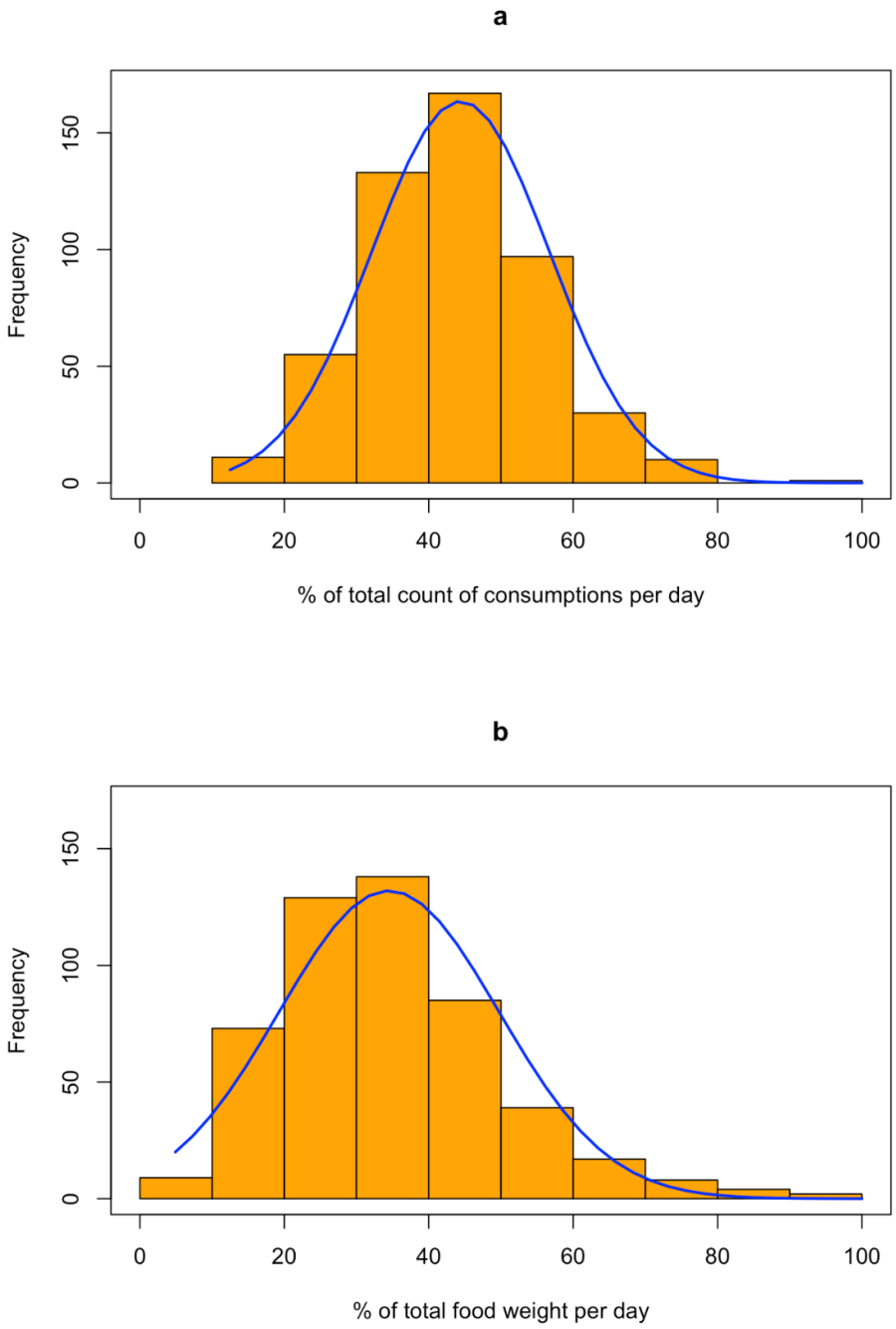
	NPNS	GUI
Sample size (n)	126	9,793
Subject age	3 years	3 years
Nationally representative	Yes	Yes
Date of survey	Oct 2010-Sept 2011	Dec 2010-July 2011
Food measurement instrument	4-day weighed food diary	Short frequency questionnaire

Table 5.1 Comparison of survey characteristics of National Preschool Nutritional Survey (NPNS) and Growing Up in Ireland (GUI) national infant cohort survey.

The frequency and amount of food consumed that was not mapped by the GUI survey is depicted in Figures 5.1a and 5.1b, respectively.

The histograms represent the distribution of the ratio of consumption counts (Figure 5.1a) or amount (Figure 5.1b) of food items consumed per person per day in NPNS that were not covered by the mapped GUI database divided by the total number of consumptions or amount, respectively, per day. For example, the ratio of consumption counts is the number of food consumptions *non-covered* by the mapped GUI model divided by the total number of food consumptions in any given day. The overall pattern of the distribution of percentage consumption frequency was symmetrical while the shape of the distribution for percentage food amount was skewed slightly to the right. The mean (SD) for consumption frequency was 44% (12%) and for consumption amount was 34% (15%). As some food codes in NPNS were partially mapped by GUI the % of coverage was estimated for all foods. For example, other fruit in NPNS was partially mapped to GUI and approximately 63% of this food group was *non-covered*.

A selection of the most commonly consumed *non-covered* (by GUI) food items during the NPNS 4-day period is displayed in Table 5.2. Food items rich in sugar that were *non-covered* included RTEBC, fruit juice, sugars, syrups, preserves and sweeteners and ice-cream.



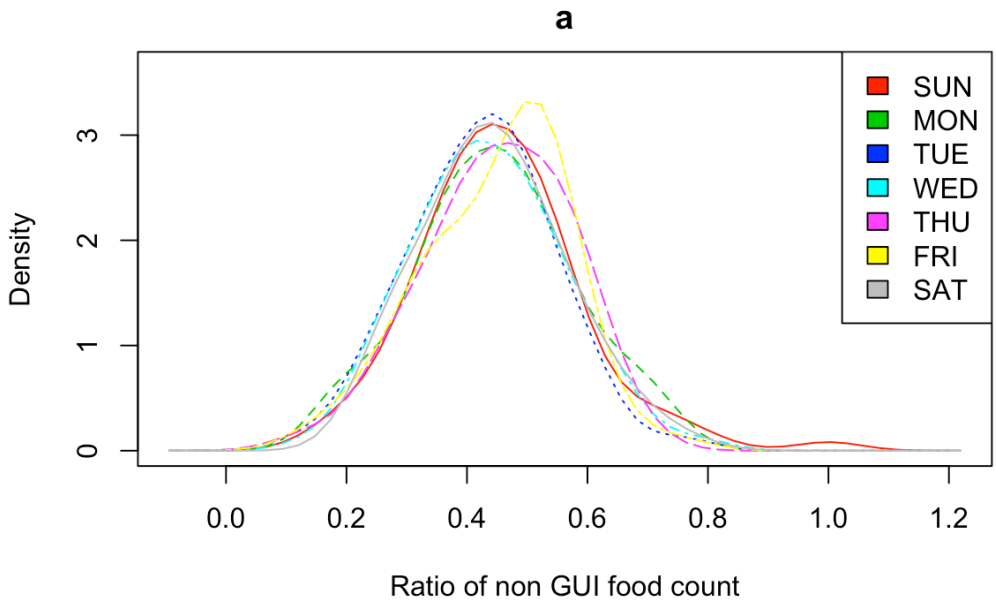
Figures 5.1a and 5.1b Food frequency and consumption weight *non-covered* by GUI survey representing the distribution of the ratio of consumption counts* (Figure 5.2a) or weight (Figure 5.2b) of a food item consumed in NPNS that were *non-covered* by the mapped GUI data model. *number of food consumptions *non-covered* by the mapped GUI model divided by the total number of foods consumed in a given day.

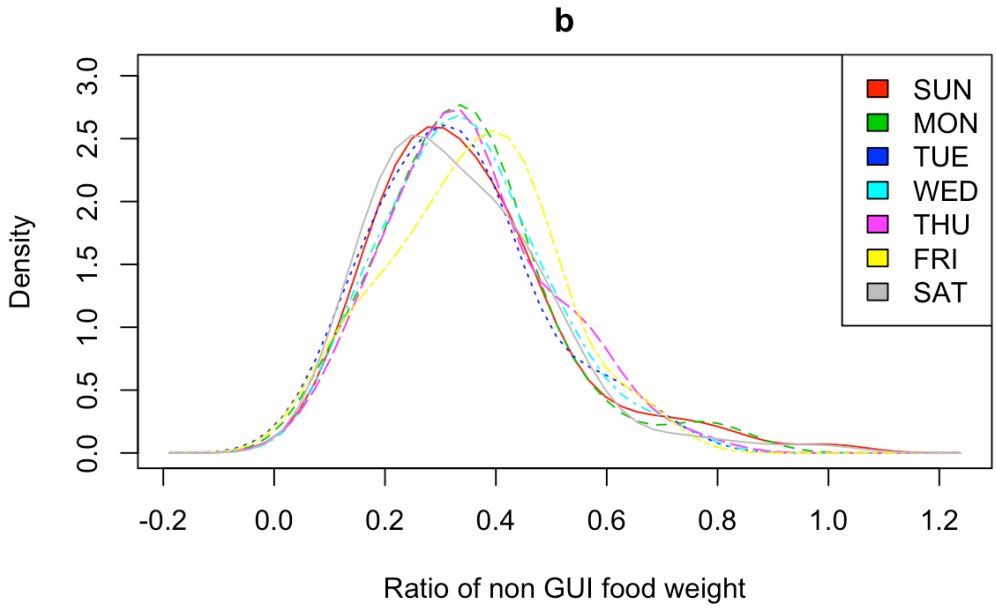
Table 5.2 Number of Eating Occasions (EO), Food amount (g/day) and Standard Deviation (SD) of selected *non-covered* food items in augmented food database.

Food group description	EO Number	Food wt (g/d) Mean	SD
RTEBC	351	31	17
White sliced bread and rolls	239	61	34
Other spreading fats	224	8	5
Wholemeal and brown bread and rolls	192	50	29
Fruit juices	190	173	123
Soups, sauces and miscellaneous foods	190	52	71
Potatoes	163	79	46
Sugars, syrups, preserves and sweeteners	158	12	13
Bacon and ham	148	30	25
Rice and pasta, flours, grains and starch	131	86	57
Supplements	124	106	60
Meat products	95	52	43
Butter	90	9	8

Ice creams	89	57	25
Beef and veal dishes	88	129	82
Chicken, turkey and game	87	44	26
Other breakfast cereals	83	130	83
Eggs and egg dishes	64	60	30
Fish and fish products	62	62	32

The distribution of the proportion of *non-covered* food (frequency of consumption and amount, g/d) by the day of the week is displayed in Figure 5.2a and Figure 5.2b as density estimates





Figures 5.2a and 5.2b Food frequency and consumption weight *non-covered* by GUI survey by the day of the week representing the distribution of the ratio of consumption counts (Figure 5.2a) or weight (Figure 5.2b) of a food item consumed in NPNS that were *non-covered* by the mapped GUI data model over the total food *covered*.

The distribution of the ratio of *non-covered* food to total food varied according to the day of the week. The distribution patterns on Friday and Sunday appeared to have some differences from the other days with a shift of the distribution to the right for Friday and an increased "tail" on Sunday. Permutation tests were carried out which omitted one day from each test while retaining all the others which suggested that the distributions for each day of the week were significantly different from each other, $p<0.01$ (except Monday for consumption frequency and Sunday and Monday for consumption amount).

5.4. Discussion

Substantial progress has been made in assessing and interpreting dietary intake data (Taren et al., 2006, Satija et al., 2015). However, as emphasised in a recent systematic review (Golley et al., 2017) there is a need to provide guidance on which questions to utilise to measure children's food intake and this will depend on the research focus and study sample. The aim of this analysis was to develop a mapping procedure that allowed for detailed dietary data from a matched cohort to be mapped to simple data from a large cohort with the aim of improving the quality of dietary data in large cohorts and therefore improve the capacity to identify diet-disease relationships. In doing so, it was possible to evaluate the performance of a SFQ compared to the "gold standard" FD for estimating food and nutrient intakes.

Rather than report the average weight or frequency of food consumed the proportion of these metrics as a percentage of the total food consumed was estimated to illustrate how much of the foods from the detailed NPNS were *covered* or *non-covered* by the SFQ used in the GUI survey. As illustrated in Figures 5.1a and 5.1b, there was a wide spread of the distribution with the average consumption frequency of foods not *covered* by GUI ranging from 22-56% while average consumption amount of *non-covered* foods by GUI ranged from 19- 49%. Thus, the SFQ in GUI did not capture a substantial portion of habitual food consumption (as estimated by a 4-day weighed food diary in a similar population) by 3 year olds in Ireland. When evaluating the relative validity of any dietary assessment tool it is important that the test method and reference method measure the same underlying concept over the same time period (Golley et al., 2017). The approach here was to use the reference method (4-day weighed FD) to map the food groups in the test method (SFQ).

The SFQ used in GUI was not designed to capture habitual food intake but reflected what is often used in large scale-interdisciplinary surveys (Thompson and Subar, 2013).

While it is thought that questions in a SFQ are more likely to be reliable than accurate (Golley et al., 2017) there is less detailed knowledge of the type of measurement error with SFQs than other more detailed instruments (Rutishauser, 2007). While SFQs are obviously appealing to include in a survey to measure dietary intake these brief screeners tend to be widely used without relative validation (Golley et al., 2017, Rutishauser, 2007, Kirkpatrick et al., 2014). The lack of accurate estimates of dietary intake may lead to biased determinations of relationships between for example, consumption of "unhealthy" food and weight status or sugar consumption and dental caries. Other researchers have highlighted the benefits of combining information from multiple surveys to gain and augment estimates of parameters lacking in individual surveys (Schenker and Raghunathan, 2007, Slack-Smith, 2012).

While this data analysis was carried out retrospectively these results highlight the importance of selecting the most appropriate dietary assessment instrument given the study design, resources and objectives. However, the protocol described here could be applied in other scenarios, particularly *post-hoc* interdisciplinary studies, to link datasets for further analysis. Where knowledge of habitual dietary intake is required it may be possible to plan the alignment of a national cohort with a similar population sample nutritional survey to maximise the value of data extraction. The use of data linkage and other techniques such as integrated health modules within longitudinal surveys should be explored.

Inappropriate feeding patterns of "unhealthy" or "sugar rich" food and drink appear to start as young as 6 months of age (Johnson et al., 2016, Amezdroz et al., 2015, Chaffee et al., 2015) and tend to increase as the child moves to solid foods in the first few years of life (Johansson et al., 2010). It would appear reasonable to use a SFQ focussed on capturing "unhealthy" food intake in a national cohort survey. However, results presented here highlight the lack of capture of some foods and drinks rich in sugar (Table 5.2) and commonly implicated in causing dental caries. As analyses of large cohort child surveys are commonly used to inform key public health policy-related issues such as oral health or childhood obesity services it is important that appropriate dietary information can be extracted to maximise the full potential of these studies. As

well as the lack of appropriate questions in the SFQ to capture these items, day-to-day variation can also contribute to insufficient estimation particularly as habitual intake of food and drinks rich in added sugar has been reported to be higher at weekends compared to weekdays (Svensson et al., 2014). In our analysis, some differences were noted in the distribution of both amount and frequency of consumption of *non-covered* food on Friday and Sunday compared to other days of the week (Figures 5.3a and 5.3b) but most days of the week showed significant differences using the permutation test.

Although the results highlight key shortfalls of the GUI SFQ, it is important to acknowledge that the GUI survey was not designed to report detailed dietary intakes per se but to use a brief screener-type SFQ which collapsed food groups into what was considered “healthy” and “unhealthy”. The categorisation could potentially introduce bias as PCGs may under or over report due to social desirability of what are perceived as “healthy” and “unhealthy” foods. Compared to other food mapping algorithms such as free sugar estimation (Louie et al., 2015) the mapping protocol in this analysis contained a low risk of subjectivity as the degree of detail included (e.g., cooking method and detailed food description) facilitated accurate mapping to match the GUI food codes.

5.5. Conclusions

This data analysis protocol provided a method for further mapping of national cohort surveys and food databases for other age cohorts. Through mapping the food codes in this manner and estimating the degree of *non-covered* food it was possible to visualise the relative performance of the brief dietary instrument compared to the more detailed one especially in capturing specific food types, e.g., high sugar foods. The SFQ did not capture a substantial portion of habitual foods consumed by 3 year olds in Ireland. Researchers interested in focussing on specific foods, such as those high in sugar, could use this approach to easily assess the proportion of foods *covered*, *non-covered* or *partially-covered* by reference to the mapped food database.

Chapter 6. Patterns of selected cariogenic food and drink intake in preschool children: linking data from two national surveys

6.1. Introduction

The importance of nutrition in infancy and childhood is well established. Early-life dietary interventions can reverse growth faltering and improve cognitive development (Bhutta et al., 2013). Despite the evidence that describes the impact of nutritional interventions during critical periods (Bhutta et al., 2013), few seem to realise that choice of foods and drinks in infancy and childhood can also influence the initiation, progression and reversal of early caries lesions (Selwitz et al., 2007, Lingström, 2009, Moynihan and Petersen, 2004). This is significant given the fact that ECC, the 10th most common chronic disease of childhood (Kassebaum et al., 2015), is associated with general health and wellbeing and impacts significantly on quality of life (Gussy et al., 2006, Johnson et al., 2016). Inappropriate consumption of cariogenic food and drink (CF) is a key determinant of ECC (Chankanka et al., 2015, Marshall et al., 2005, Chaffee et al., 2015, Tinanoff and Reisine, 2009, Dye et al., 2004, Marshall et al., 2007b). The carious process occurs through the metabolism of fermentable dietary carbohydrates at the plaque-biofilm interface resulting in localised demineralisation and destruction of enamel and dentine over time (Selwitz et al., 2007, Bradshaw and Lynch, 2013). Therefore, foods and drinks such as confectionary, carbonated beverages, squashes, cordials, fruit juice drinks, cakes and biscuits are considered potentially cariogenic (Moynihan, 2002). Excessive intake of these food groups is also associated with overweight and obesity-related disorders (Hooley et al., 2012a, Marshall et al., 2007b, Amine et al., 2002).

Meal frequency and the amount of sugars or CF are strongly related (Moynihan and Petersen, 2004) but there is still debate about which plays a more dominant role in dental caries progression (Diaz-Garrido et al., 2016, Moynihan and Kelly, 2014). Recently, Diaz-Garrido and co-workers (2016) examined, *in vitro*, the effect of increased frequency of sucrose exposure on the cariogenicity of a biofilm-caries model and concluded that even one daily exposure to sugar can initiate a carious lesion. Additionally, dietary studies have shown that children with a history of snacking once or more per day or children who consume sweets a minimum of once a day have higher rates of dental caries (Abreu et al., 2015). However, while we know there is a strong relationship between CF intake and dental caries in young children (Chaffee et al., 2015, Dye et al., 2004, Johansson et al., 2010, Marshall et al., 2005, Llena and Forner, 2008) the importance of the pattern of intake of CF (frequency x amount) for ECC has not been fully elucidated (Burt et al., 1988, Sheiham, 2007, Chankanka et al., 2015, Johnson et al., 2016). Given the changing and dynamic nature of dietary habits in preschool children it is important to develop a better understanding of the pattern of CF intake; for example, are CF foods typically consumed as components of main meals or snacks, and are they consumed with other CF items? Also, food consumption surveys typically report mean daily intakes for nutrients and foods rather than amounts per eating occasion (EO). Establishing what the threshold is for the number of "exposures" to CF per day or the amount per EO required to induce ECC would have implications for health education and caries prevention strategies. The diet of Irish preschoolers generally meets recommendations for most nutrients (Walton et al., 2017); however, a better understanding of dietary patterns is needed so that strategies can be developed to reduce the intake of sugar containing food and drink in this age group.

Current dental professional recommendations are that the first dental visit should occur at 12 months of age, primarily to offer preventive advice to parents regarding appropriate oral health behaviours (American Academy of Pediatric Dentistry, 2013). Dentists are required to provide dietary advice to reduce the frequency and amount of potentially CF and promote good oral health (Moynihan, 2002). However, this practice is not standardised and there is no dietary assessment system to monitor patient diets or CF intake. Therefore, there is a need to develop a practical, valid, and reliable dietary assessment tool for use in the clinical dental setting (Arheiam et al., 2016a, Arheiam et al., 2016b).

Even in a research setting, comparisons across studies are difficult as there is no standardised method for reporting dietary intake data (Faber et al., 2013, Moynihan et al., 2009). Studies often use food frequency questionnaires (FFQ) or short food questionnaires (SFQ) to reduce respondent burden, despite limitations with regards to food quantification (Llena and Forner, 2008, Zero, 2004, Johansson et al., 2010). Another limitation when examining diet and dental caries is that most studies have focused on nutrient-disease interactions. However, people eat combinations of food items as snacks or main meals rather than individual nutrients (Leech et al., 2015) and few studies have specifically examined patterns of consumption of CF in preschool children.

The aim of the present study was to investigate the pattern of consumption of selected CF in 3 year old preschool children using the NPNS food database (Irish Universities Nutrition Alliance, 2012) mapped with food intake codes for 3 year olds from the GUI survey; This novel mapped dataset was then analysed to explore the pattern of dietary intake of selected CF items as snacks or main meals. The mapped dataset allowed the effect of using a SFQ on dietary intake data to be visualised, compared to a more detailed weighed dietary assessment. This study adds to previous reporting of the risks involved when selecting SFQs for large national cohort studies which may be less costly and burdensome than more detailed methods but may not sufficiently capture the quality of data required to investigate diet related health outcomes. The analysis makes detailed comparisons of different methods to report and understand the pattern of CF consumption.

6.2. Methods

The aim in this chapter was to investigate the pattern of consumption of selected cariogenic food and drink (CF) in 3 year old preschool children using the NPNS food database mapped with food intake codes for 3 year olds from the GUI survey (section 2.6.3). This mapped dataset was then analysed to explore the pattern of dietary intake of selected CF items as snacks or main meals. The items were preselected according to their potential cariogenicity as indicated in Table 1.3, Section 1.5.1 (Moynihan, 2002). Not all sources of fermentable carbohydrate were included and foods such as yogurt (with added sugar), sweetened milks and starch products were omitted. The mean daily intake amount (g/day) and the frequency of each food consumed was

calculated for each NPNS participant, by summing the amount of all foods a subject consumed per food code, averaging across the four days for each subject and then calculating the total sample average. Frequency was estimated by summing the total number of times the food appeared in the diary and dividing by four, i.e. the number of days in the survey. The average consumption amount (g/day) and frequency for each food was computed by aggregating the data across each subject and each survey day before averaging the food weight. Thus, if a food was consumed more than once on a given day the average consumption amount was calculated to provide a closer representation of the actual amount of food consumed on a single EO. If a food was not consumed at all on a given day, it was not included in the estimations for average consumption amount. All results reported were for consumers only.

Global statistics were generated for the number of subjects who either never consumed a CF or consumed a CF. Each meal type (both 'snacks' and 'meals') was defined by its food components. A food component was defined as a single food item from each meal. The number of food components, both CF and non-cariogenic food and drink (NCF), in a meal, were determined at the meal level and at the subject level (Chapter 2.6.6).

Bean plots were generated in R Studio for dietary intake estimated using both mean daily intake and average consumption methods. Asymmetric bean plots allowed an easy comparison between two subgroups such as main meals and snacks and revealed anomalies such as bimodal distributions (Kampstra, 2008). Main meals were coloured blue while snacks were coloured red. CF items (rice puddings and custard, tinned fruit and carbonated beverages) with a small number of subjects (<20) were not included. A two-sample Kolmogorov-Smirnov test ($p<0.05$ alfa level) was carried out to determine if the distributions of CF items as a snack or main meal, as estimated by both mean daily intake and average consumption, were similar. Further details are provided in Section 2.6.6.

Association analysis was carried out using the NPNS database mapped with the GUI food codes to: (1) identify the food components that comprised a meal; (2) compare meals by chaining the components within a meal using either the GUI coding or the NPNS (n=77) Food group coding; (3) identify the combinations of food components that characterise the most frequently consumed meals and (4) assess the dietary interaction of each CF with NCF and the other CF items selected. Variables included were: subject ID, meal

type, time of consumption, NPNS 77 FG, food weight and GUI food code. An association analysis algorithm generated frequent item sets (Williams, 2011). Association analysis is a useful methodology for discovering interesting relationships hidden in a dataset. Further details of the novel association analysis can be found in Section 2.6.7 and Appendix.

6.3. Results

6.3.1. Cariogenic food and beverage intake

A comparison of the survey characteristics of GUI and NPNS is presented in Table 6.1. All subjects consumed at least one CF item during the NPNS 4-day period. From a total of 2,676 meals, 1,500 (56%) contained at least one CF item. Only one subject did not consume snacks and the average frequency of snacks consumed by each subject was 2 per day. Approximately 36% of "all eating occasions" were categorised as snacks and 76% of all subjects consumed 2 or more snacks per day. The average frequency of consumption of all CF was 3.9, calculated by mean daily intake, or 4.3, calculated by average consumption. The estimates, using mean daily intake, of selected CF by total, snacks and main meals for consumers only are presented in Table 6.2.

Table 6.1 Comparison of survey characteristics of National Preschool Nutritional Survey (NPNS) and Growing Up in Ireland (GUI) Longitudinal study of children-infant cohort.

	NPNS	**GUI**
Sample size (n)	500 (126=3 years old)	9,793 (wave 2=3 years old)
Study type	Cross- sectional	Longitudinal
Nationally representative	Yes	Yes
Date of survey	Oct 2010-Sept 2011	Dec 2010-July 2011
Food measurement instrument	4-day weighed food diary	Modified 24-hour FFQ

Table 6.2 Cariogenic food eating occasions [frequency and amount (Food wt, g/d) estimated using mean daily intake method for consumers only.

		All Eating Occasions					Main Meal					Snacks				
			Food wt (g/d)		Frequency		Food wt (g/d)			Frequency			Food wt (g/d)		Frequency	
NPNS Food group	*GUI Code*	*n*	*Mean*	*SD*	*Mean*	*SD*	*n*	*Mean*	*SD*	*Mean*	*SD*	*n*	*Mean*	*SD*	*Mean*	*SD*
RTEBC	NC	116	24	14	0.9	0.5	114	23	13	0.9	0.5	19	8	5	0.3	0.1
Biscuits including crackers	C25g	95	15	12	0.9	0.8	48	7	5	0.5	0.5	88	13	10	0.7	0.5
Cakes, pastries and buns	C25g	55	13	11	0.4	0.2	33	11	9	0.3	0.2	36	10	9	0.3	0.2
Ice creams	NC	59	22	14	0.4	0.3	34	18	12	0.4	0.2	37	17	10	0.3	0.1
Desserts	C25g	23	25	22	0.4	0.2	16	22	22	0.4	0.2	12	18	19	0.3	0.1
Rice puddings and custards	NC	15	41	43	0.4	0.3	11	32	33	0.3	0.2	7	37	24	0.3	0.1
Fruit juices	NC	78	105	81	0.8	0.5	70	89	65	0.7	0.4	31	64	51	0.4	0.2

Tinned fruit	NC	5	14	14	0.3	0.1	4	17	15	0.3	0.1	1	3	NA	0.2	NA
Sugars, syrups, preserves and sweeteners	NC	71	6	7	0.7	0.5	62	6	6	0.7	0.4	26	4	3	0.4	0.2
Chocolate confectionery	C25g	75	11	8	0.5	0.3	27	8	5	0.4	0.1	62	9	7	0.4	0.2
Non-chocolate confectionery	C25h	57	10	7	0.5	0.3	25	8	5	0.4	0.2	43	9	7	0.4	0.3
Carbonated beverages (non-diet)	C25m	20	69	52	0.4	0.4	15	66	45	0.4	0.3	8	49	23	0.3	0.2
Squashes, cordials and fruit juice drinks	C25l	84	84	93	1.2	0.9	74	61	69	1.0	0.7	62	40	45	0.5	0.3

NPNS is the National Preschool Nutritional Survey; GUI is Growing Up in Ireland; n= number of subjects; NC= *non-covered* by GUI food codes; Eating Occasion (EO)= All 'snacks' and 'main meals' were collectively described as 'eating occasions'.

Table 6.2 (continued) Cariogenic food eating occasions [frequency and amount (Food wt, g/d) estimated using mean daily intake method for consumers only

Of the 13 CF items selected from the augmented database, six were not covered by GUI food codes. The largest contributors to CF intake in the NPNS sample in terms of mean frequency and mean amount (g) were RTEBC, biscuits, fruit juices and squashes cordials/ fruit juice drinks.

Using mean daily intake approach for calculating intakes, the mean frequency (SD) of individual CF items ranged from 0.3 (0.1) for tinned fruit to 1.2 (0.9) for squashes, cordials and fruit juice drinks. Mean food weight (g/d, SD) varied from 6 (7) for sugars, syrups, preserves and sweeteners to 105 (81) for fruit juices. RTEBC was consumed by almost all of the sample population while biscuits including crackers, fruit juices and squashes, cordials and fruit juice drinks were consumed by approximately 75%, 62% and 67% of all the sample population, respectively. Chocolate confectionary and non-chocolate confectionary were consumed at a similar mean frequency with both items approximately twice as likely to be consumed as a snack rather than as part of a main meal.

Carbonated beverages (non-diet) were consumed by 16% of the total sample. The mean frequency intake of RTEBC as a main meal was three times greater than when consumed as a snack and was largely consumed as a main meal by most individuals. Biscuits including crackers were almost twice more likely to be consumed as a snack and the mean frequency intake was also greater as a snack compared to consumption as part of a main meal.

The estimates, using average consumption, of selected CF by total, snacks and main meals for consumers only are presented in Table 6.3.

Table 6.3 Cariogenic food eating occasions [frequency and amount (Food wt, g/d)] estimated using average consumption method for consumers only.

		All Eating Occasions					**Main Meal**					**Snacks**				
			Food wt (g/d)		**Frequency**			**Food wt (g/d)**		**Frequency**			**Food wt (g/d)**		**Frequency**	
NPNS Food group	*GUI code*	*n*	*Mean*	*SD*	*Mean*	*SD*	*n*	*Mean*	*SD*	*Mean*	*SD*	*n*	*Mean*	*SD*	*Mean*	*SD*
RTEBC	NC	116	26	10	1.2	0.4	114	26	10	1.2	0.4	19	26	11	1.1	0.2
Biscuits including crackers	C25g	95	18	9	1.4	0.8	48	16	10	1.3	1.0	88	19	10	1.4	0.8
Cakes, pastries and buns	C25g	55	33	19	1.1	0.4	33	33	19	1.1	0.2	36	32	17	1.0	0.2

Ice creams	NC	59	54	24	1.1	0.2	34	54	29	1.1	0.2	37	55	19	1.0	0.2
Desserts	C25g	23	65	55	1.1	0.3	16	57	31	1.1	0.3	12	68	73	1.1	0.3
Rice puddings and custards	NC	15	95	55	1.0	0.2	11	96	60	1.0	0.1	7	114	70	1.0	0.0
Fruit juices	NC	78	141	68	1.3	0.4	70	133	65	1.2	0.3	31	150	85	1.1	0.2
Tinned fruit	NC	5	55	59	1.0	0.0	4	66	62	1.0	0.0	1	12	NA	1.0	NA
Sugars, syrups, preserves and sweeteners	NC	71	10	9	1.2	0.4	62	9	9	1.2	0.3	26	10	6	1.1	0.2
Chocolate confectionery	C25g	75	21	9	1.3	0.7	27	22	13	1.1	0.3	62	21	10	1.2	0.6
Non-chocolate confectionery	C25h	57	24	18	1.2	0.4	25	22	15	1.1	0.3	43	24	19	1.2	0.4

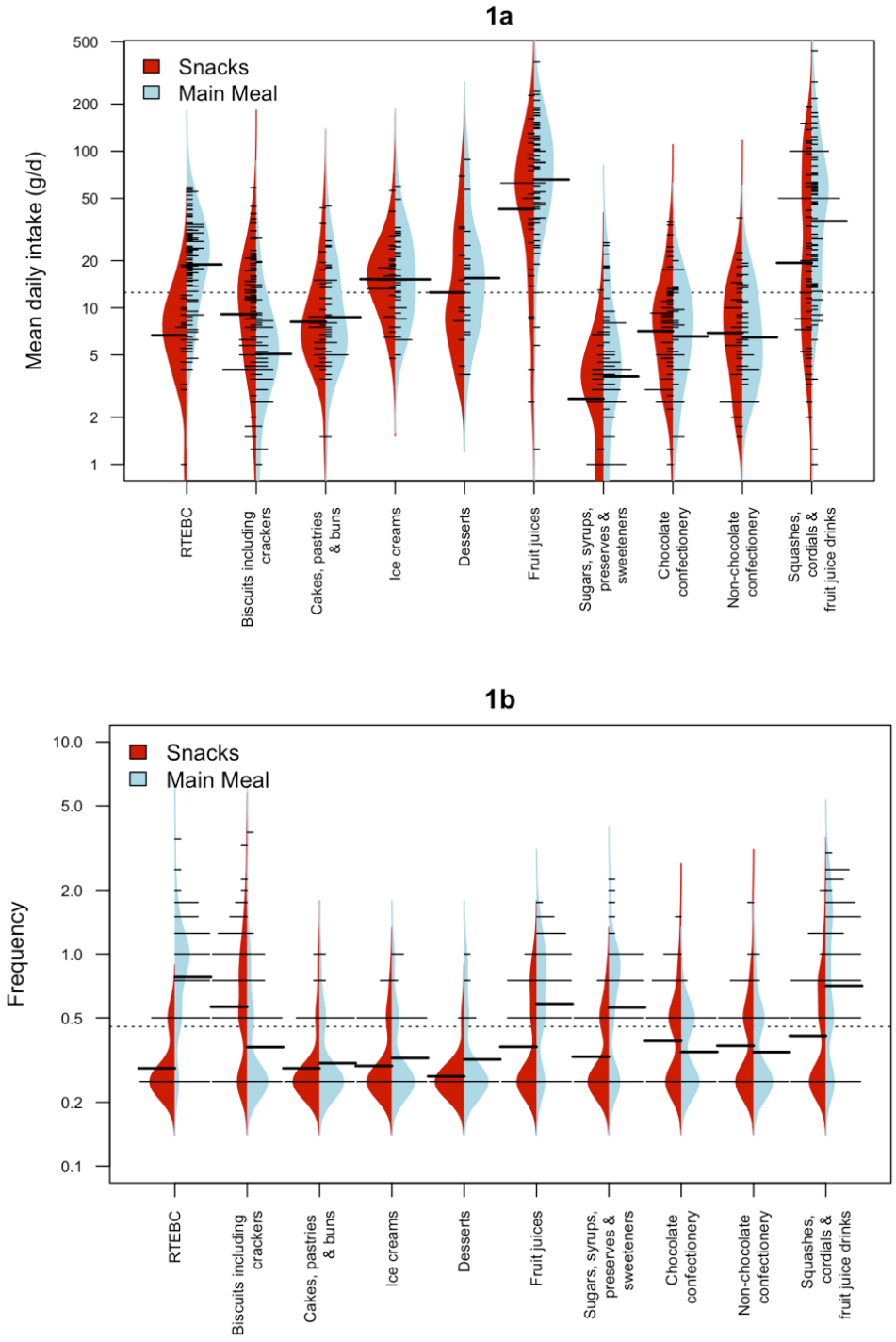
Carbonated beverages (non-diet)	C25m	20	174	50	1.2	0.5	15	168	56	1.1	0.4	8	163	46	1.0	0.0
Squashes, cordials and fruit juice drinks	C25l	84	80	73	1.7	0.8	74	77	70	1.5	0.6	62	81	79	1.1	0.2

NPNS is the National Preschool Nutritional Survey; GUI is Growing Up in Ireland; n= number of subjects; NC= *non-covered* by GUI food codes; Eating Occasion (EO)= All 'snacks' and 'main meals' were collectively described as 'eating occasions'.

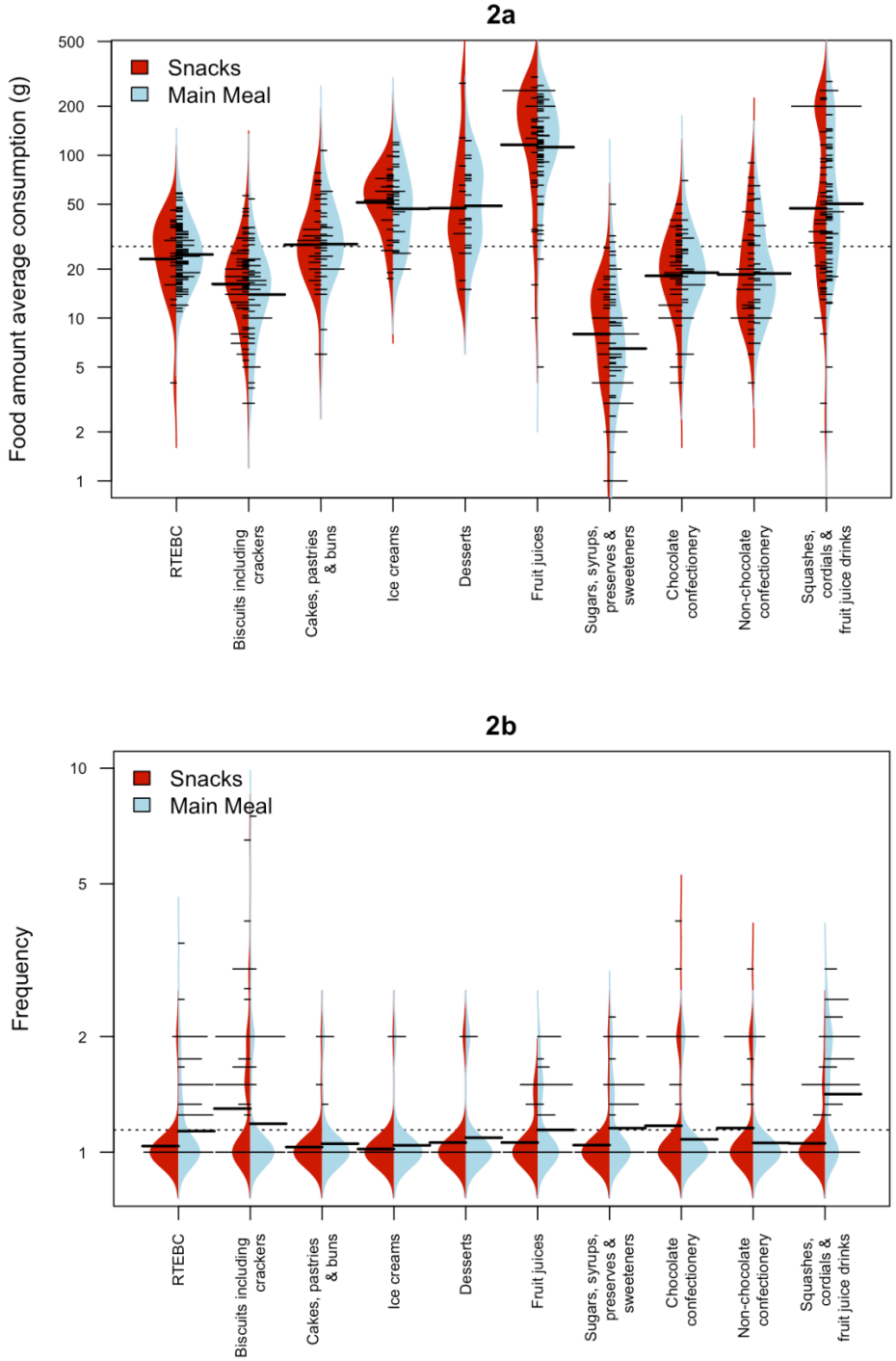
Table 6.3 (continued from previous page) Cariogenic food eating occasions [frequency and amount (Food wt, g/d)] estimated using average consumption method for consumers only

Total mean frequency intake (mean, SD) of CF was highest for squashes, cordials and fruit juice drinks (1.7, 0.8) followed by biscuits including crackers (1.4, 0.8) and fruit juices (1.3, 0.4). The highest total mean weight intakes (g/d, SD) of cariogenic drinks were carbonated beverages (174, 50) and fruit juices (141, 68). The highest mean weight intakes (g/d, SD) of cariogenic foods were rice puddings and custards (95, 55) and desserts (65, 55).

When main meals and snacks were compared, the items with the highest mean frequency were biscuits including crackers as a snack (1.4, 0.8) and squashes, cordials and fruit juice drinks as part of a main meal (1.5, 0.6). The Kolmogorov-Smirnov tests suggested no significant difference ($p<0.05$) between the two distributions for both frequency and amount for snacks compared to main meals when calculated using average consumption. However, there was a significant difference for almost half of the CF items when the Kolmogorov-Smirnov test was applied to compare the distributions for snacks and main meals using mean daily intake estimates. Bean plots (Figures 6.1.a, 6.1.b, 6.2.a, and 6.2.b) illustrated the kernel density estimates of these consumption measures and allowed comparison of differences in overall distribution patterns of main meals and snacks using the two methods of calculating intakes. For example, as illustrated in Figure 6.1, using mean daily intake, RTEBC were consumed at a lower frequency as a snack and a higher frequency as a main meal with a bimodal distribution for snack consumption and a lower mean frequency. However, while the mean frequency for RTEBC consumption as a meal is higher the spread of the distribution is much wider. Using the average consumption method, the different shapes of bean plots for snacks and main meals of the distribution profile for frequency of consumption of squashes, cordials and fruit juice drinks were easily visualised in Figure 6.2b while the plots for cakes, pastries and buns were similar.



Figures 6.1a and 6.1b Bean plots illustrating the distribution patterns (kernel density estimates) of weight of food consumed (Figure 6.1a) and frequency of consumption (Figure 6.1b) of snacks and main meals estimated using the mean daily intake method. The shape of the bean plots highlights the differences in distribution of snacks (red) and main meals (blue) as a subgroup of each CF item. The mean for each subgroup is indicated by the heavy solid horizontal line.



Figures 6.2a and 6.2b. Bean plots illustrating the distribution patterns (kernel density estimates) of weight of food consumed (Figure 6.2.a) and frequency of consumption (Figure 6.2.b) of snacks and main meals estimated using the average consumption method. The shape of the bean plots highlights the differences in distribution of snacks (red) and main meals (blue) as a subgroup of each CF item. The mean for each subgroup is indicated by the heavy sold horizontal line

6.3.2. Association analysis of eating occasions

The association analysis aimed to identify the components of all EO and identify the combinations of components in the most commonly consumed meals and snacks. Using the keys for NPNS-derived food codes the 2,676 total meal observations were collapsed to 1,622 nodes of the tree and then using the keys for GUI-derived food codes the observations were collapsed to 456. Aggregating the meal observations into CF and NCF items using the keys for NPNS-derived food code collapsed the observations to 163. On average, each meal was composed of 3.44 food components, 25% of these components were CF and there were 0.81 cariogenic components per meal. Nearly all subjects had a meal that contained at least 1 or 2 CF components during the survey period. Table 6.4. presents the top 20 most commonly consumed meals using keys for GUI-derived food codes.

Table 6.4 Association analysis, using keys* for GUI†-derived food codes, showing the number of subjects and food amount for the most commonly consumed main meal or snack and total number of eating occasions (EO) *covered* and *non-covered* by GUI food codes.

KEY*	GUI food	Meal Type	EO	n	Food wt (g/d)	GUI *covered* Food wt (g/d)	GUI *non-covered* Food wt (g/d)
C25n-NC	Full cream milk/milk products	main	214	72	237	142	95
NC		main	136	73	182	NA	182
C25a	Fresh fruit	snack	125	54	106	106	NA
C25g	Biscuits, doughnuts cake, pie or chocolate	snack	89	56	28	28	NA
NC		snack	81	55	95	NA	95
C25k-NC	Water	main	72	48	280	117	163
C25n	Full cream milk/milk products	main	60	25	191	191	NA
C25i-NC	Full fat cheese/yoghurt/fromage frais	main	52	35	245	76	169
C25o-NC	Skimmed/Semi-skimmed	main	44	20	228	139	89
C25k-C25l-NC	Water-Fizzy drinks (diet)	main	42	23	361	188	173
C25a-C25n-NC	Fresh fruit-Full cream milk or milk products	main	41	24	333	214	119
C25g-NC	Biscuits, doughnuts cake, pie or chocolate	snack	41	36	107	25	83
C25a-NC	Fresh fruit	main	39	30	260	79	181
C25i	Full fat cheese/yoghurt/fromage frais	snack	37	27	102	102	NA
C25b-C25k-NC	Cooked vegetables-Water	main	36	27	356	216	140
C25b-NC	Cooked vegetables	main	32	25	221	40	181
C25a-C25k	Fresh fruit-Water	snack	30	19	239	239	NA
C25g-NC	Biscuits, doughnuts cake, pie or chocolate	main	30	26	201	46	155
C25b-C25n-NC	Cooked vegetables-Full cream milk/milk products	main	29	20	311	189	122

C25i-C25n-NC	Full fat cheese/yoghurt/fromage frais-Full cream milk or milk products	main	29	24	298	217	81

* Keys for GUI-derived food codes were were derived by association analysis of all eating occasions and restructuring the data into a tree shape. A set of distinct GUI food codes were generated and foods *not-covered* by GUI were labelled as NC to build a unique identifier for each meal. This meal identifier was the key for GUI-derived food codes.

†Growing up In Ireland; n= number of subjects; EO= All 'snacks' and 'main meals' were collectively described as 'eating occasions'; NC= *non-covered* by GUI food code; NA= not-applicable.

Table 6.4 (continued) Association analysis, using keys* for GUI†-derived food codes, showing the number of subjects and food amount for the most commonly consumed main meal or snack and total number of eating occasions (EO) *covered* and *non-covered* by GUI food codes.

Of these, 6 were 'snacks' and 14 were 'main meals'. Of the 10 most commonly consumed meals, 3 were fully *covered*, 2 were *non-covered* and the remainder were *partially covered* in the GUI survey by GUI food codes. Fresh fruit was the most frequently consumed snack over the 4-day survey period and was eaten by 43% of the NPNS population followed by biscuits, doughnuts, cakes, pie or chocolate which was consumed by 44% of the total NPNS sample.

The second association analysis carried out used the set of NPNS (FG77) food group codes (n=77) to build a unique identifier for each meal. This allowed a comparison between meal components described using the GUI food codes and meal components described using NPNS food codes. As well as assessing how well the SFQ in GUI performed in capturing cariogenic food components in meals or snacks this also illustrated how reporting of consumption of foods was affected by how 'coarse' (GUI) or 'fine-grained' (NPNS) the food codes were. Table 6.5 presents the top 20 most commonly consumed meals using keys for NPNS-derived food codes. Of the 10 most commonly consumed meals, as defined using NPNS-derived keys, 8 were snacks. The most commonly consumed snack was other fruit (except bananas which were coded separately) followed by chocolate confectionary and biscuits including crackers which were consumed by 34, 24 and 21%, respectively, of the total NPNS sample.

The association analysis of CF items for the 20 most frequently consumed meals containing CF is presented in Table 6.6. Most of these meals were composed of only one or two CF items which were consumed more often as part of a main meal. Using these descriptors (cariogenic and non-cariogenic aggregates), the most commonly consumed CF items as a snack were biscuits, including crackers, squashes, cordials and fruit juice drinks and chocolate confectionary which constituted approximately 17%, 29% and 100% respectively, of the snacks. RTEBC, squashes, cordials and fruit juice drinks and fruit juices were components of three of the most commonly consumed main meals with fruit juice constituting approximately 48% of the total meal amount (Table 6.6). Overall, biscuits including crackers featured as a component in four of the twenty most frequently consumed meals and three of those eating occasions were snacks.

Table 6.5 Association analysis, using keys* for NPNS† -derived food codes, showing the most commonly consumed main meal or snack, total number of eating occasions (EO), number of subjects who consumed the meal and weight of food.

	KEY	Meal Type	Total number EO	Total subject count	Average Food wt (g)	SD Food wt (g)
1	Other fruit	snack	70	43	90	59
2	RTEBC-Whole milk	main	62	36	172	71
3	Whole milk	main	42	20	190	85
4	Chocolate confectionery	snack	38	30	24	11
5	Biscuits including crackers	snack	37	26	24	12
6	Yogurts	snack	23	17	119	52
7	Non-chocolate confectionery	snack	22	20	26	22
8	Savoury snacks	snack	20	17	23	13
9	Bananas	snack	19	16	110	54
10	Biscuits including crackers-Whole milk	snack	19	8	175	59
11	Ice creams	snack	18	16	57	21
12	Other milks and milk based beverages	main	18	5	194	100
13	RTEBC-Whole milk-Sugars, syrups, preserves and sweeteners	main	18	11	188	76
14	Other beverages	main	17	12	107	70
15	Other beverages-Squashes, cordials and fruit juice drinks	main	17	10	200	61

16	RTEBC-Low fat, skimmed and fortified milks	main	17	7	176	85
17	Other fruit-Other beverages	snack	14	11	240	145
18	Citrus fruits	snack	12	6	98	70
19	Whole milk	snack	12	7	199	64
20	Yogurts	main	12	7	110	38

*Keys for NPNS -derived food codes were derived by association analysis of all eating occasions and restructuring the data into a tree shape. The tree was analysed using a set of distinct NPNS 77 Food group codes; †National Preschool Nutrition Survey

EO= All 'snacks' and 'main meals' were collectively described as 'eating occasions'.

Table 6.5 (continued) Association analysis, using keys* for NPNS† -derived food codes, showing the most commonly consumed main meal or snack, total number of eating occasions (EO), number of subjects who consumed the meal and weight of food.

Table 6.6 Association analysis of 20 most frequent eating occasions using keys* for NPNS†-derived food codes for cariogenic and non-cariogenic descriptors.

Rank	Key Descriptor	Meal Type	Frequency	Subject count	RTEBC	Biscuits including crackers	Ice-creams	Desserts	Rice puddings and custard	Fruit juices and smoothies	Tinned fruits	Sugars, syrups, preserves and sweeteners	Chocolate confectionary	Non-chocolate confectionary	Carbonated beverages	Squashes, cordials and fruit juice drinks	NCF
1	RTEBC-NCF	main	187	85	15	0	0	0	0	0	0	0	0	0	0	0	85
2	Squashes, cordials and fruit juice drinks-NCF	main	172	57	0	0	0	0	0	0	0	0	0	0	0	20	80
3	Biscuits including crackers-NCF	snack	97	53	0	17	0	0	0	0	0	0	0	0	0	0	83
4	Fruit juices-NCF	main	80	45	0	0	0	0	0	48	0	0	0	0	0	0	52
5	Squashes, cordials and fruit juice drinks-NCF	snack	62	43	0	0	0	0	0	0	0	0	0	0	0	29	71
6	Sugars, syrups, preserves and sweeteners-NCF	main	48	29	0	0	0	0	0	0	0	5	0	0	0	0	95

7	RTEBC-Fruit juices-NCF	main	46	26	11	0	0	0	0	37	0	0	0	0	0	0	52
8	RTEBC-Sugars, syrups, preserves and sweeteners-NCF	main	45	22	14	0	0	0	0	0	0	3	0	0	0	0	83
9	RTEBC-Squashes, cordials and fruit juice drinks-NCF	main	44	23	8	0	0	0	0	0	0	0	0	0	0	25	67
10	Chocolate confectionery	snack	38	30	0	0	0	0	0	0	0	0	100	0	0	0	0
11	Biscuits including crackers	snack	37	26	0	100	0	0	0	0	0	0	0	0	0	0	0
12	Chocolate confectionery-NCF	snack	27	22	0	0	0	0	0	0	0	0	25	0	0	0	75
13	Biscuits including crackers-NCF	main	26	23	0	9	0	0	0	0	0	0	0	0	0	0	91
14	Ice creams-NCF	main	23	19	0	0	23	0	0	0	0	0	0	0	0	0	77
15	Fruit juices-NCF	snack	22	18	0	0	0	0	0	51	0	0	0	0	0	0	49
16	Non-chocolate confectionery	snack	22	20	0	0	0	0	0	0	0	0	0	100	0	0	0
17	Cakes, pastries and buns-NCF	main	22	18	0	0	0	0	0	0	0	0	0	0	0	0	86
18	Biscuits including crackers-Squashes, cordials and fruit juice drinks-NCF	snack	21	16	0	15	0	0	0	0	0	0	0	0	0	18	67

19	Fruit juices-Sugars, syrups, preserves and sweeteners-NCF	main	19	12	0	0	0	0	0	38	0	4	0	0	0	0	58
20	Non-chocolate confectionery-NCF	snack	19	15	0	0	0	0	0	0	0	0	0	18	0	0	82

*Keys for NPNS -derived food codes were derived by association analysis of all eating occasions and restructuring the data into a tree shape. The tree was analysed using a set of distinct NPNS 77 Food group codes; †National Preschool Nutrition Survey;

EO= All 'snacks' and 'main meals' were collectively described as 'eating occasions'; NCF= Non-cariogenic food.

Table 6.6 (continued) Association analysis of 20 most frequent eating occasions using keys* for NPNS†-derived food codes for cariogenic and non-cariogenic descriptors.

Focussing on cariogenic food interactions, alluvial plots were generated to visualise the component interactions between the 10 most frequent eating occasions of cariogenic food and presented in Figure 6.3. However, data for the specific CF components in the alluvial also included any other food group interactions (outside the top 10) with CF and NCF. This resulted in all other interactions outside of the most frequent ones to be included in the alluvial. An alluvial diagram is a form of parallel coordinates plot used for categorical variables. Similar to parallel coordinates, variables are positioned on vertical axes that are parallel. The blocks on each axis represent values, e.g., snack or main meal in the first axis of Figure 6.3, and the height of a block represents the size of the values. The coloured 'streams' (alluvials) connecting the blocks are called flows (blue for main meals and red for snacks) and the width of the flows illustrates the proportion or frequency in each category (Rosvall and Bergstrom, 2010). For example, starting on the left hand side of the alluvial in Figure 6.3 and following the blue flow for biscuits (including crackers), the blue flow connecting the lowest point of the main meal block to the first cariogenic component (cg1) (biscuits) then connects to 'none' on the second cariogenic component (cg2) indicating that there is no other cariogenic component in that meal. The blue flow from the bottom of the 'none' block at cg2 then goes to the '90' block on the NCF (non-cariogenic food) axis and, finally, connects to the first block (biscuits –NCF) at the meal components axis. The red flow for the same meal in this block can be followed in reverse also and this indicates that biscuits (including crackers) when consumed as a snack were consumed with a high proportion of NCF (85) and mainly without a second CF component (cg2). The pattern of interactions for each of the CF groups can be traced in a similar manner as a snack or main meal.

To visualise a simpler set of interactions, a second alluvial plot is shown in Figure 6.4 illustrating association analysis of 2 eating occasions with cariogenic food using keys[*] for NPNS[†]-derived food codes for cariogenic and non-cariogenic descriptors.

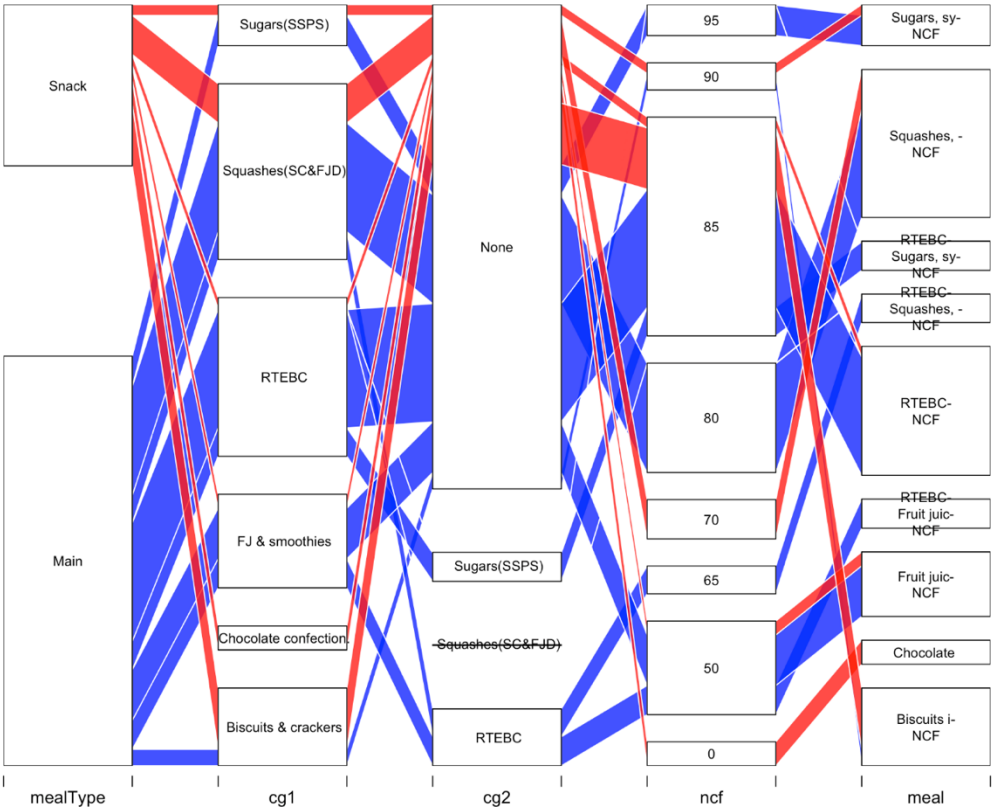


Figure 6.3 Alluvial plot of association analysis of 10 most frequent eating occasions using keys* for NPNS†-derived food codes for cariogenic and non-cariogenic descriptors

*Keys for NPNS -derived food codes were derived by association analysis of all eating occasions and restructuring the data into a tree shape. The tree was analysed using a set of distinct NPNS 77 Food group codes; †National Preschool Nutrition Survey; Blue flow= main meals; Red flow= snacks; cg1= first cariogenic food component; cg2= second cariogenic food component; ncf= non-cariogenic food component; NCF= non-cariogenic food; Sugars (SSPS)= sugars, syrups, preserves and sweeteners; FJ= fruit juice; RTEBC= ready to eat breakfast cereals.

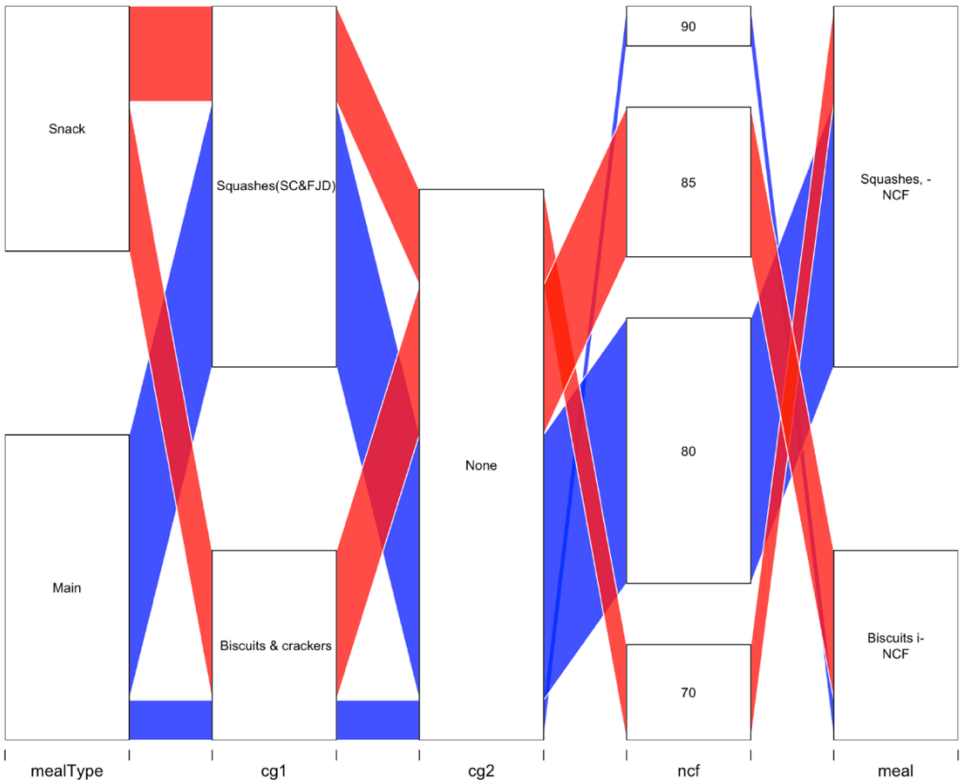


Figure 6.4 Alluvial plot illustrating association analysis of 2 eating occasions with cariogenic food using keys* for NPNS†-derived food codes for cariogenic and non-cariogenic descriptors.

*Keys for NPNS -derived food codes were derived by association analysis of all eating occasions and restructuring the data into a tree shape. The tree was analysed using a set of distinct NPNS 77 Food group codes; †National Preschool Nutrition Survey; Blue flow= main meals; Red flow= snacks; cg1= first cariogenic food component; cg2= second cariogenic food component; ncf= non-cariogenic food component; NCF= non-cariogenic food; SC&FJD= squashes, cordials and fruit juice drinks.

6.4. Discussion

Comparing consumer intake of CF using mean daily intake and consumption average (Tables 6.2 and 6.3, Figures 6.1a, 6.1b, 6.2a and 6.2b) suggested that the method of reporting data can influence interpretation and understanding of CF intake patterns. While it is useful to know the mean daily intake of nutrients, the focus of this study was to estimate consumer average intake of typical meal or snack portion sizes and frequency to understand the differences in CF distribution. This approach could help provide practical information to consumers and health professionals attempting to modify the risk of dental caries and other diet-related health problems (Leech et al., 2015, Moynihan, 2002, Krebs-Smith et al., 2015, Bhutta et al., Amine et al., 2002). Overall intake of CF was high, whether measured by mean daily intake or average consumption, compared to recommended dietary guidelines for children (Food Safety Authority of Ireland, 2011b, Food Safety Authority of Ireland, 2011a). From an oral health perspective, there was sufficient evidence to justify a focus on 'snacking habits' to help minimise both frequency and amount of exposure to CF (Moynihan and Petersen, 2004, Dye et al., 2004, Moynihan, 2002, Moynihan and Kelly, 2014, Meyer and Lee, 2015, Sheiham and James, 2015, Johansson et al., 2010). Results here suggested that the distribution pattern was similar for snacks and main meals when using the average consumption method; there was little variation in the frequency (except for squashes, cordials and fruit juice drinks) or food amount distribution across all CF items. This was illustrated in the similar shapes of the kernel density in the asymmetric bean plots (Figures 6.2a, 6.2b) and the subgroup average for each CF item as depicted by the closely approximating solid horizontal lines. Thus, in terms of actual exposure to CF both main meals and snacks contributed similarly to frequency and amount of intake.

This study also aimed to determine the extent to which CF items are not covered by a simple SFQ used in a large national cohort survey and explore the pattern of intake of CF using data from the NPNS survey. Findings indicated that all children consumed CF over the NPNS 4-day period and the GUI survey covered less than half of the CF items selected in NPNS. More than one-third of all EO were described as 'snacks' which were consumed twice per day, on average. Meal analysis using keys for GUI-derived food codes suggested that it achieved only partial coverage of most of the top ten most commonly consumed meals and snacks.

Fruit was the most commonly consumed snack (using either keys for GUI or NPNS-derived food codes) followed by chocolate confectionary (key for NPNS-derived food codes) or biscuits, doughnuts, cakes, pie or chocolate (key for GUI-derived food codes). Results presented in Section 4 using the GUI infant cohort data indicated that the majority of Irish children aged 3 years consumed fresh fruit (~89%), and biscuits/doughnuts/cake/chocolate (~74%) once or more than once per day. Results of the meal association analysis were strongly dependant on how "coarse" or "fine-grained" the food categories and food codes used in each survey are. The modified SFQ used in GUI survey did not quantify food intake but did capture almost half of the selected CF items. However, for investigating CF and NCF the meal association analysis using GUI-derived keys highlighted the "coarseness" of the GUI food categories. The association analysis using the NPNS-derived keys was too "fine-grained" but the association analysis tree was further re-shaped using NPNS food descriptors to define all CF and aggregating the NCF together. This permitted increased focus on the interactions between CF and NCF in a meal analysis (Table 6.6) which suggested that for the top twenty most frequently consumed meals CF items are generally consumed with NCF items, as a main meal or snack, except for three meal types where chocolate confectionary, biscuits including crackers and non-chocolate confectionary were consumed without NCF as a snack. The alluvial diagram (Figure 6.3) visualised the trend of CF consumption as a snack or main meal and if consumed with another cariogenic component or not. Following the red flow pattern CF snacks were visualised and it was apparent that sugars, syrups, preserves and sweeteners (SSPS), chocolate confectionary and biscuits including crackers were mainly consumed as snacks and without a second cariogenic component. While squashes, cordials and fruit juice drinks (SC&FJD) were consumed as both a snack and part of a main meal, mostly without a second cariogenic component but a small portion was consumed with RTEBC and a substantial non-cariogenic food (milk). The second alluvial plot presented in Figure 6.3 displayed association analysis of 2 eating occasions with cariogenic food to provide a simple example of the interaction patterns. This higlights the differences in snacking patterns between biscuits including crackers and squashes, cordials and fruit juice drinks.

Most of the CF items selected in this study contained negligible nutrient content. However, while fruit juices, fruit juice drinks and RTEBC contain relatively high levels of free sugar it has also been suggested that they can provide an

important source of nutrients for young children with RTEBC also usually consumed with milk (Priebe and McMonagle, 2016, Moynihan and Petersen, 2004, Moynihan, 2002, Dye et al., 2004). While most studies in children have not found an association between 100% fruit juice consumption and ECC it is generally advisable to limit the intake for overall health reasons as the high sugar content can contribute to increased calorie consumption (Heyman and Abrams, 2017, Gussy et al., 2006, Marshall et al., 2005). Fruit juices and squashes, cordials and fruit juice drinks are, potentially, both cariogenic and erosive due to their high sugar content and relative acidity (Marshall et al., 2003, Marshall, 2013, Gussy et al., 2006). In our analysis, fruit (non-dried), which is generally regarded as non-cariogenic, was the most common 'snack' food in between main meals and it has been suggested that increasing consumption of whole fruit rather than fruit juice or other CF items would be a healthier alternative for young children that could also reduce the risk of dental caries (Moynihan and Petersen, 2004, Dye et al., 2004, Heyman and Abrams, 2017). While the cariogenic potential of any fermentable carbohydrate depends on the way it is consumed, evidence suggests that regular soft drink consumption is more strongly associated with ECC than intake of 100% fruit juice (Marshall et al., 2003, Johnson et al., 2016, Chankanka et al., 2015, Tinanoff and Palmer, 2000). However, at this early age, non-diet carbonated beverages were consumed by only 16% of the NPNS sample whereas fruit juice was consumed by 62% of the sample and 'squashes, cordials and fruit juice drinks' consumed by 67% of the sample. Interestingly, fruit juices and carbonated beverages (non-diet) were, approximately, twice more likely to be consumed as part of a main meal than as a snack while squashes, cordials and fruit juice drinks were consumed by a similar number of the NPNS sample whether as a snack or main meal component. Packaged and pre-sweetened RTEBC's are an important source of nutrients (Priebe and McMonagle, 2016) and potentially cariogenic (Moynihan, 2002) but, as demonstrated in this analysis (Table 6.6, Figure 6.3) and reported previously, are mainly consumed with milk products as part of a main meal which may reduce their cariogenic effects (Dye et al., 2004). Clearly, the manner and pattern of consumption of these type of CF items that contribute significant nutrients requires further investigation but should be considered when assessing the overall potential cariogenicity and acidogenicity of a child's diet (Moynihan, 2002, Arheiam et al., 2016a).

Dietary counselling for parents or caregivers targeting CF has been suggested as a key preventive strategy for ECC (Johnson et al., 2016, Hooley et al., 2012c, Selwitz et al., 2007, Johansson et al., 2010, Marshall et al., 2005, Tinanoff and Palmer, 2000, Moynihan and Petersen, 2004). A high intake of CF, particularly as snacks, has been associated with an increased risk of ECC in multiple studies (Marshall et al., 2005, Chaffee et al., 2015, Johansson et al., 2010, Dye et al., 2004, Chankanka et al., 2015). The results from this analysis showed that exposure to potentially CF items was widespread in 3-year old Irish preschoolers with an average consumption frequency of four times per day Guidelines suggesting an upper threshold for sugar intake, while useful at a population level, have questionable practical benefit to the individual consumer or patient in a medical or dental setting (Arheiam et al., 2016a). While it is well established that all infants have an innate liking for sweetness in food and drink the introduction of CF is strongly influenced by the PCG (Bonotto et al., 2017). Rather than adopt a "zero-tolerance" for all sugar containing food and drink it may be more practical to focus on those CF items consumed as snacks that have minimal nutritional content to encourage healthy eating practices. Targeting the reduction of CF snacks and limiting CF consumption to main mealtimes only, has long been a recommendation for the prevention of dental caries (Moynihan, 2002, Moynihan and Petersen, 2004). Current recommendations by the WHO emphasise the importance of reducing both frequency and amount (<10% energy intake) of free sugar containing food and drink to reduce the risk of dental caries (Moynihan, 2016, Moynihan and Kelly, 2014). However, it is important to be aware that these are population guidelines and the translation of these recommendations requires multiple practical strategies (Moynihan, 2016). This analysis focussed on the most commonly consumed discretionary CF items, illustrated the importance of understanding portion sizes and highlights the need to standardise methods of interpreting and reporting food intake data (Faber et al., 2013). The results also highlight the usefulness of visualising the overall distribution of intake when comparing consumption patterns (Figures 6.1a, 6.1b, 6.2a and 6.2b).

Data used in this secondary analysis was derived from two nationally representative surveys. The dietary intake data from NPNS provided detail at the brand level (Irish Universities Nutrition Alliance, 2012) and a trained nutritionist visited the participant during the 4-day survey period. The pattern of intake of food was analysed using two different methods to estimate consumption of individual food items and an association analysis of the meals

and snacks provided an insight into the meal components and how CF was consumed with other NCF. While the NPNS included detailed food records the GUI survey was not designed to look at dietary intakes per se and therefore condensed food groups into what they considered “healthy” and “unhealthy” to reduce respondent burden. The categorisation could potentially introduce bias as PCGs may under or over report due to social desirability of what is perceived as ‘healthy’ and ‘unhealthy’ foods. Our analysis treated each day of the 4-day period in NPNS as an independent day. The association analysis assumed that all meals were independent. Neither NPNS nor GUI included a clinical dental examination so there was not an opportunity to examine possible associations with an index of dental caries. The distinction between a ‘meal’ and ‘snack’ was also subjectively decided by the PCG. It is also worth noting that this study did not include all sources of fermentable carbohydrate as CF items, e.g. yogurt, sweetened milks and starch products. The true cariogenicity of any food can only be determined by exposing humans to the food and measuring the associated tooth decay (Lingström, 2009, Lingstrom et al., 2000). However, there is sufficient evidence to suggest that refined starch is potentially cariogenic especially in combination with sucrose (Lingstrom et al., 2000, Lingström, 2009, Johansson et al., 2010).

The preschool age is a particularly important period to minimise the risks for dental caries as it is the best predictor of future dental caries (Gussy et al., 2016, Tinanoff and Palmer, 2000, Dye et al., 2015, Peres et al., 2016). It also appears to be a key transition phase when targeting preventive strategies to focus on appropriate intake of ‘healthy’ food items and restriction of CF items could provide the best opportunity for non-cavitated caries lesions to arrest (Johnson et al., 2016, Chaffee et al., 2015, Amezdroz et al., 2015). However, given the differing terminology for sugars and lack of information about free or added sugars both on food labels and in nutrient databases, it is difficult to see how consumers and health care providers can assess compliance with sugar intake recommendations (Erickson and Slavin, 2015b). Sugar is rarely ingested in a pure form (Zero, 2004) and this analysis demonstrated the complexity of meal analysis when CF items are consumed with both other ingredients and other food components in meals having the potential to modify the cariogenicity of the food components (Tinanoff and Palmer, 2000, Lingström, 2009, Marshall et al., 2005, Moynihan and Petersen, 2004). Even if RTEBC, fruit juice and tinned fruit are excluded from the preselected CF items, as they are not included in the “top shelf” of the food pyramid, it is clear that a majority of three-

year old preschool children in Ireland do not meet the guidelines for healthy eating and consume CF items every day and, on average, more than three times per day rather than a “maximum once or twice a week” (Food Safety Authority of Ireland, 2011b, Food Safety Authority of Ireland, 2011a). Rather than focus on a threshold for sugar intake it is, arguably, more practical to concentrate on those CF items, particularly consumed as ‘snacks’, that do not contribute significantly to nutrient intake but are primary contributors to CF intake. For three-year olds this would include biscuits including crackers, squashes, cordials, fruit juice drinks and chocolate confectionary. While carbonated beverages (non-diet) were not widely consumed by the NPNS sample (16% consumers) they do not contribute significant nutrients. Reducing or substituting the intake of these CF items with healthier alternatives, as snacks, could have an immediate benefit in reducing the frequency and amount of free sugar intake.

This analysis, supported by recent studies (Johnson et al., 2016, Goodwin et al., 2017, Arheiam et al., 2016a, Arheiam et al., 2016b) suggests that while food diaries are labour intensive for both respondent and analyst the potential benefits to provide dietary advice from both a population and personalised patient perspective should be acknowledged and developed further. There is an immediate need to establish universal definitions for ‘snacking’ so that study participants and researchers are clear on what constitutes a snack (Jacquier et al., 2017, Leech et al., 2015) and to provide appropriate training for health care professionals on dietary assessment and healthy eating strategies. Investigating food related factors that may affect cariogenicity, other than frequency and amount, such as food texture effects on mastication and oral clearance, could open up useful future research strategies for caries prevention.

6.5. Conclusions

Dietary intake data from smaller studies can augment the information collected in larger national surveys where detailed food diaries may not be employed. This analysis demonstrated not only the importance of selecting the most appropriate dietary instrument to achieve survey objectives but also that the methods by which data is analysed and reported affects the interpretation of the results. Our results suggest that the overall intake of CF was high compared

to recommended dietary guidelines for children. A better understanding of the distribution pattern of CF intake provides an opportunity for appropriate dietary intervention of cariogenic snacking by focussing on foods and meals which may significantly contribute to the risk of ECC in preschoolers.

Chapter 7. Estimation and consumption pattern of free sugar intake in 3 year old pre schoolers

7.1. Introduction

Dietary sugars are the focus of debate in public health in recent years with concerns regarding increased obesity prevalence and their impact on oral health (Allison et al., 2015, SACN, 2015, Pyne and Macdonald, 2016). As dietary free sugars (FS) are the most important risk factor in the development of dental caries and can contribute to excess energy intake with little nutrient benefit, the WHO have issued recommendations to apply over the lifetime; firstly, that adults and children reduce their daily intake of FS to <10% of Total Energy Intake (TEI) (strong recommendation) and secondly, that a further reduction to <5% of TEI would provide additional health benefits (conditional recommendation) (World Health Organization, 2015).

As outlined in Section 1.5, the term “sugar” refers to sucrose or “table sugar” (Marshall, 2015). Total sugars (TS) is the sum of natural and added sugars (AS) in a food, while intrinsic sugars can be defined as the sugars incorporated in the structure of intact fruit and vegetables (Marshall, 2015) (Figure 1.4). AS include those sugars added during the production or processing of food and not naturally found in the food product (Erickson and Slavin, 2015b) and is the term defined by the Food and Drug Administration (FDA) in the USA. Free sugars (FS), which is the preferred term used by the WHO, includes sugars naturally present in honey, syrups, fruit juices and fruit juice concentrates as well as AS. AS and FS are not chemically distinguishable from those sugars naturally occurring in food and drink. This has important consequences for difficulties in determining the FS content of foods, food labelling, consumer understanding and compliance with new guidelines on FS consumption. Understanding the problems in measuring FS and assessing the difficulties for consumers in meeting the recommended guidelines is important to make sugar

reduction targets achievable for both the food industry in formulation (Public Health England, 2017) and consumers in practical terms (Erickson and Slavin, 2015b, Erickson and Slavin, 2015a). Recommendations by the Scientific Advisory Committee on Nutrition (SACN) were that the average population intake of FS should not be greater than 5% of TEI from 2 years upwards and that for children there should be a minimal consumption of sugar-sweetened drinks (SACN, 2015).

Worldwide, there are very few food databases that contain information regarding AS or FS levels (Moynihan et al., 2018, Erickson and Slavin, 2015b, Newens and Walton, 2016, Brisbois et al., 2014). This may be due to difficulties in estimating AS or FS levels (Louie et al., 2015, Newens and Walton, 2016) and the lack of agreed standards in terms of definitions, terminology, standardisation of estimation procedures and protocols for dealing with the food industry in terms of ingredients/recipe and sugar reduction programmes (Public Health England, 2017, Erickson and Slavin, 2015a, Erickson and Slavin, 2015b). As outlined in Section 1.6, a number of recent reviews have examined TS/AS/FS consumption measured by dietary surveys worldwide (Newens and Walton, 2016, Azais-Braesco et al., 2017) and concluded that further research is urgently required to address the deficit of information and measures to address the problem of excess intake. In most national studies, a majority, or at least a large proportion, of children exceeded the recommended guidelines for AS/FS consumption (Lei et al., 2016, Ruiz et al., 2017, Gibson et al., 2016, Brisbois et al., 2014, Azais-Braesco et al., 2017, Farajian et al., 2016), and, consequently, some professional paediatric bodies have made practical recommendations for reducing AS/FS intake in children (Mis et al., 2017).

The preschool age is a key period when cariogenic food and drink is already part of the child's diet. Both progression and reversal of early caries can occur, and sugar intake is a key risk factor, yet preschool children have rarely been included in most national oral health surveys (Mis et al., 2017, Amezdroz et al., 2015). The WHO expert consultation on public health intervention against early childhood caries have recommended the inclusion of 3 year old age group as one of the index ages for population surveys in the next edition of WHO's *Oral health surveys: basic methods* (World Health Organization, 2017b). Also, in the first survey of 3 year old children in England approximately 12% were reported to have had dental caries (Public Health England, 2013a). While there are no nationally representative data in Ireland for dental caries prevalence in 3 year olds, some smaller scale studies have reported dental caries prevalence rates

of 0-41.5% (O'Connell and Harding, 2017) (Table 1.2). Our objectives were to: (1) quantify FS using the previously mapped NPNS and GUI food data (chapter 5); (2) determine the distribution of TS and FS consumption patterns (amount and frequency) in relation to the WHO guidelines; (3) compare how well the SFQ in GUI captured the sources of TS/FS and (4) identify the key food sources of TS/FS consumed as part of a main meal or snack.

7.2. Methods

7.2.1. Data source

Data used for this analysis was derived from 3 year olds in the GUI and NPNS studies as outlined in Chapter 2.6.

7.2.2. Data work flow

Methodology used in the initial stages of data entry, wrangling and estimation of *covered* and *non-covered* GUI food data and NPNS food data has been detailed in Section 2.6.1-2.6.5. A flow chart presented in Figure 7.1. outlines the workflow used in the mapping and analysis of FS.

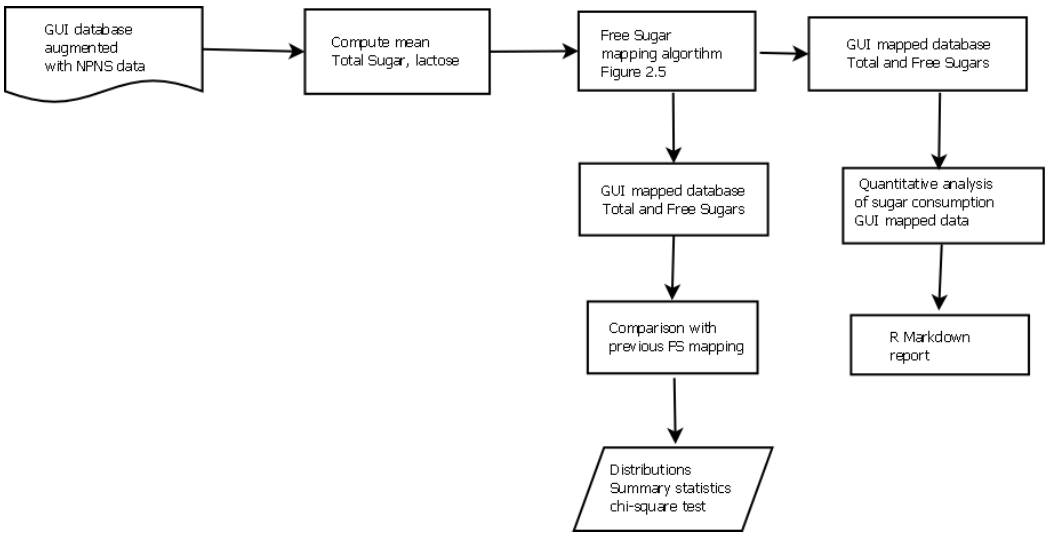


Figure 7.1 Flow chart depicting the processes in the mapping and analysis of free sugar intake using the Growing Up in Ireland (GUI) database augmented with the national preschool nutrition survey (NPNS).

This diagram, created using Dia (http://dia-installer.de/), represents the processes and steps taken as various types of boxes and their order is shown by using interconnecting arrows. Essentially, the mapped GUI database (Section 2.6.3) was further augmented by carrying out a mapping of FS (Section 2.7). The decision algorithm used for estimating FS content of food groups from NPNS is illustrated in Figure 2.5, Section 2.7. This FS estimation was then compared to previously published data from the same cohort (Newens and Walton, 2016) and the distributions compared using the Kolmogorov-Smirnov tests ($p<0.01$). Quantitative analysis and metrics similar to that detailed in Section 2.6.5 were carried out. All statistical analyses were carried out using R Studio (https://www.rstudio.com/). Complete details of these analyses are included in Appendix A. TS and FS were determined by multiplying the weight of food consumed daily, aggregated at the subject level, by the percentage of TS or FS. The mean daily intake of TS and FS (g/d), frequency of consumption and as a percentage of total energy intake, (TEI) were presented as summary statistics. The daily intake of TS and FS *covered* and *non-covered* by GUI food groups by amount (g/day) and as %TEI were presented as bar graphs.

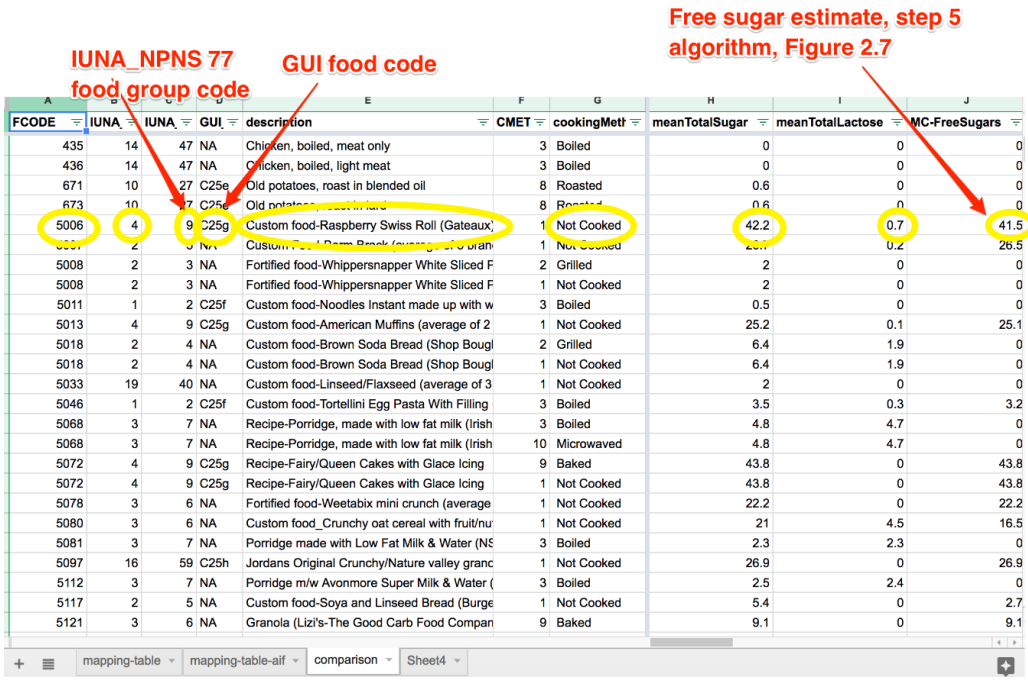
Food groups for both GUI and NPNS were also recategorised as follows to highlight the main FS food sources: bread and cereals, RTEBC, cakes and biscuits, dairy products, desserts and puddings, fruit and vegetables, fruit juice and smoothies, sugar and syrups, chocolate confectionary, non-chocolate confectionary, soft drinks (non-diet), soft drinks (diet) and other. Dairy products included all milk, yoghurt, cheese and ice-cream products. Soft drinks (non-diet) included carbonated beverages, squashes, cordials and fruit juice drinks. Breads and cereals included all rice, pasta, grains and cereal based products except RTEBC. All other food categories were grouped in to ‘other foods’ (Appendix A). A simulation was run with snack consumption of selected discretionary items “removed” to estimate the effect on % FS intake and the proportion of children who would then meet the recommended threshold of 10% TEI as FS. Data reported are for average daily intake (mean, SD) across the full sample and the percentage of consumers of each food category were included. Under-reporters were not removed from the data. The probability of consuming a food group as part of a main meal or snack was calculated using the total number of meal types (snacks and main meals) that were recorded

during the 4 day survey. Full details of the methodology are reported in Section 2.7.

7.3. Results

7.3.1. FS mapped database

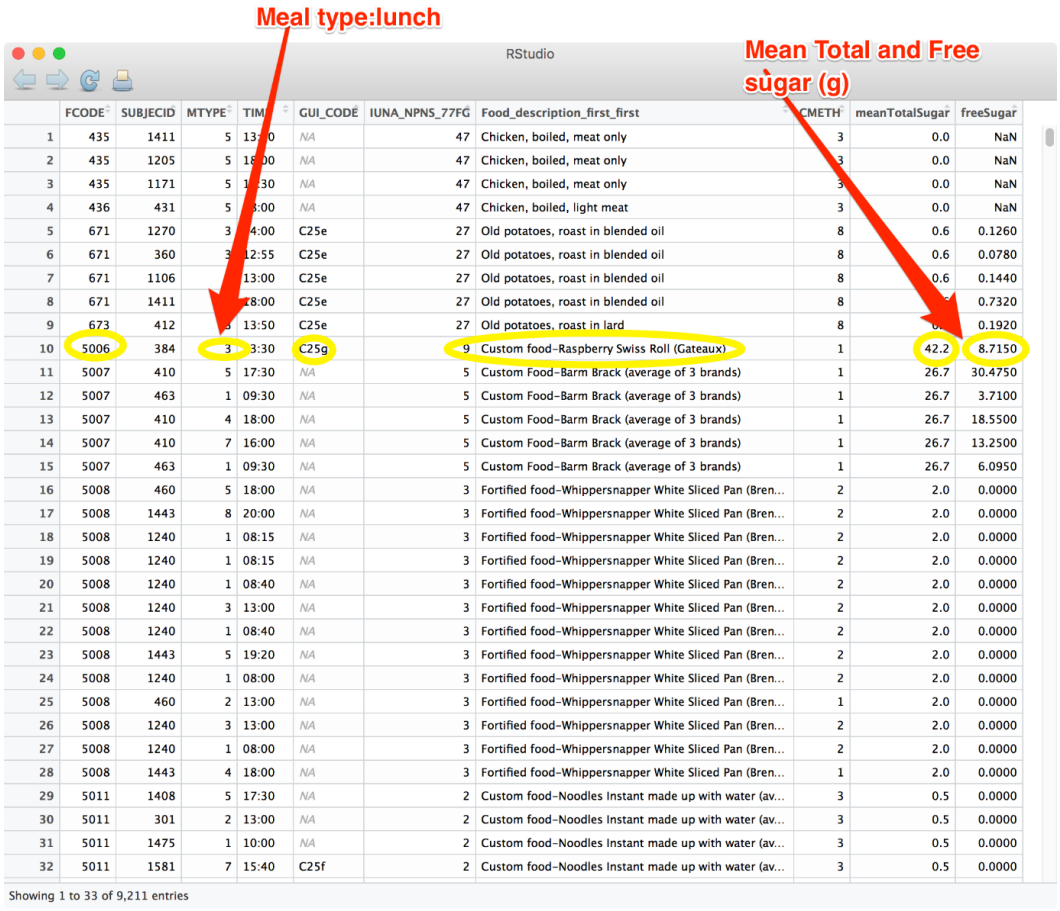
A snippet of the mapped FS database is illustrated in Figure 7.2 and the R code used to generate the input, comparison and analysis is contained in Appendix A. The resulting data was imported into RStudio and is illustrated in Figure 7.3 This data contained the food code (from a total of 1,652 food codes), the NPNS food group code (n=19) and the more detailed NPNS food group code (n=77), the GUI food code, food description, cooking method, mean total sugar, mean total lactose and estimation of FS amount according to the algorithm in Figure 2.5.



FCODE	IUNA	IUNA	GUI	description	CMET	cookingMeth	meanTotalSugar	meanTotalLactose	MC-FreeSugars
435	14	47	NA	Chicken, boiled, meat only	3	Boiled	0	0	0
436	14	47	NA	Chicken, boiled, light meat	3	Boiled	0	0	0
671	10	27	C25e	Old potatoes, roast in blended oil	8	Roasted	0.6	0	0
673	10	27	C25e	Old potatoes, roast in lard	8	Roasted	0.6	0	0
5006	4	9	C25g	Custom food-Raspberry Swiss Roll (Gateaux)	1	Not Cooked	42.2	0.7	41.5
5007	2	3	NA	Custom food-Farm Break (average of 3 bran	1	Not Cooked	26.7	0.2	26.5
5008	2	3	NA	Fortified food-Whippersnapper White Sliced F	2	Grilled	2	0	0
5008	2	3	NA	Fortified food-Whippersnapper White Sliced F	1	Not Cooked	2	0	0
5011	1	2	C25f	Custom food-Noodles Instant made up with w	3	Boiled	0.5	0	0
5013	4	9	C25g	Custom food-American Muffins (average of 2	1	Not Cooked	25.2	0.1	25.1
5018	2	4	NA	Custom food-Brown Soda Bread (Shop Bougl	2	Grilled	6.4	1.9	0
5018	2	4	NA	Custom food-Brown Soda Bread (Shop Bougl	1	Not Cooked	6.4	1.9	0
5033	19	40	NA	Custom food-Linseed/Flaxseed (average of 3	1	Not Cooked	2	0	0
5046	1	2	C25f	Custom food-Tortellini Egg Pasta With Filling	3	Boiled	3.5	0.3	3.2
5068	3	7	NA	Recipe-Porridge, made with low fat milk (Irish	3	Boiled	4.8	4.7	0
5068	3	7	NA	Recipe-Porridge, made with low fat milk (Irish	10	Microwaved	4.8	4.7	0
5072	4	9	C25g	Recipe-Fairy/Queen Cakes with Glace Icing	9	Baked	43.8	0	43.8
5072	4	9	C25g	Recipe-Fairy/Queen Cakes with Glace Icing	1	Not Cooked	43.8	0	43.8
5078	3	6	NA	Fortified food-Weetabix mini crunch (average	1	Not Cooked	22.2	0	22.2
5080	3	6	NA	Custom food_Crunchy oat cereal with fruit/nu	1	Not Cooked	21	4.5	16.5
5081	3	7	NA	Porridge made with Low Fat Milk & Water (NS	3	Boiled	2.3	2.3	0
5097	16	59	C25h	Jordans Original Crunchy/Nature valley granc	1	Not Cooked	26.9	0	26.9
5112	3	7	NA	Porridge m/w Avonmore Super Milk & Water (	3	Boiled	2.5	2.4	0
5117	2	5	NA	Custom food-Soya and Linseed Bread (Burge	1	Not Cooked	5.4	0	2.7
5121	3	6	NA	Granola (Lizi's-The Good Carb Food Compan	9	Baked	9.1	0	9.1

mapping-table | mapping-table-aif | comparison | Sheet4

Figure 7.2 Snippet illustrating the free sugar mapping procedure of the NPNS food data. FCODE: IUNA-NPNS detailed food code at level of food brand; IUNA 19: 19 food group codes; IUNA 77: 77 food group codes; GUI: food code in GUI; CM: cooking method. Example, FCODE 5006: Custom food-Raspberry Roll (Gateaux) contained (by weight) 41.5% total sugar; 0.7% total lactose and was assigned at Step 5 of decision algorithm (Figure 2.7); Total sugar minus total lactose resulted in 42.2%-0.7%= 41.5% free sugars.



	FCODE	SUBJECID	MTYPE	TIM	GUI_CODE	IUNA_NPNS_77FG	Food_description_first_first	CMETH	meanTotalSugar	freeSugar
1	435	1411	5	13:00	NA	47	Chicken, boiled, meat only	3	0.0	NaN
2	435	1205	5	18:00	NA	47	Chicken, boiled, meat only	3	0.0	NaN
3	435	1171	5	13:30	NA	47	Chicken, boiled, meat only	3	0.0	NaN
4	436	431	5	13:00	NA	47	Chicken, boiled, light meat	3	0.0	NaN
5	671	1270	3	14:00	C25e	27	Old potatoes, roast in blended oil	8	0.6	0.1260
6	671	360	3	12:55	C25e	27	Old potatoes, roast in blended oil	8	0.6	0.0780
7	671	1106	3	13:00	C25e	27	Old potatoes, roast in blended oil	8	0.6	0.1440
8	671	1411	3	18:00	C25e	27	Old potatoes, roast in blended oil	8	1.2	0.7320
9	673	412	3	13:50	C25e	27	Old potatoes, roast in lard	8	0.3	0.1920
10	5006	384	3	13:30	C25g	9	Custom food-Raspberry Swiss Roll (Gateaux)	1	42.2	8.7150
11	5007	410	5	17:30	NA	5	Custom Food-Barm Brack (average of 3 brands)	1	26.7	30.4750
12	5007	463	1	09:30	NA	5	Custom Food-Barm Brack (average of 3 brands)	1	26.7	3.7100
13	5007	410	4	18:00	NA	5	Custom Food-Barm Brack (average of 3 brands)	1	26.7	18.5500
14	5007	410	7	16:00	NA	5	Custom Food-Barm Brack (average of 3 brands)	1	26.7	13.2500
15	5007	463	1	09:30	NA	5	Custom Food-Barm Brack (average of 3 brands)	1	26.7	6.0950
16	5008	460	5	18:00	NA	3	Fortified food-Whippersnapper White Sliced Pan (Bren...	2	2.0	0.0000
17	5008	1443	8	20:00	NA	3	Fortified food-Whippersnapper White Sliced Pan (Bren...	2	2.0	0.0000
18	5008	1240	1	08:15	NA	3	Fortified food-Whippersnapper White Sliced Pan (Bren...	2	2.0	0.0000
19	5008	1240	1	08:15	NA	3	Fortified food-Whippersnapper White Sliced Pan (Bren...	2	2.0	0.0000
20	5008	1240	1	08:40	NA	3	Fortified food-Whippersnapper White Sliced Pan (Bren...	2	2.0	0.0000
21	5008	1240	3	13:00	NA	3	Fortified food-Whippersnapper White Sliced Pan (Bren...	2	2.0	0.0000
22	5008	1240	1	08:40	NA	3	Fortified food-Whippersnapper White Sliced Pan (Bren...	2	2.0	0.0000
23	5008	1443	5	19:20	NA	3	Fortified food-Whippersnapper White Sliced Pan (Bren...	2	2.0	0.0000
24	5008	1240	1	08:00	NA	3	Fortified food-Whippersnapper White Sliced Pan (Bren...	2	2.0	0.0000
25	5008	460	2	13:00	NA	3	Fortified food-Whippersnapper White Sliced Pan (Bren...	1	2.0	0.0000
26	5008	1240	3	13:00	NA	3	Fortified food-Whippersnapper White Sliced Pan (Bren...	2	2.0	0.0000
27	5008	1240	1	08:00	NA	3	Fortified food-Whippersnapper White Sliced Pan (Bren...	2	2.0	0.0000
28	5008	1443	4	18:00	NA	3	Fortified food-Whippersnapper White Sliced Pan (Bren...	1	2.0	0.0000
29	5011	1408	5	17:30	NA	2	Custom food-Noodles Instant made up with water (av...	3	0.5	0.0000
30	5011	301	2	13:00	NA	2	Custom food-Noodles Instant made up with water (av...	3	0.5	0.0000
31	5011	1475	1	10:00	NA	2	Custom food-Noodles Instant made up with water (av...	3	0.5	0.0000
32	5011	1581	7	15:40	C25f	2	Custom food-Noodles Instant made up with water (av...	3	0.5	0.0000

Showing 1 to 33 of 9,211 entries

Figure 7.3 Sample of dataframe in RStudio, illustrating the mapped GUI database (from NPNS) with mean total sugar content of food items (g/100g) and amount of free sugar in the meal (g).

This FS mapping was then compared to one previously reported for the same cohort with the summary statistics presented in Table 7.1 and the distributions presented in Figures 7.4a and 7.4b. The Kolmogorov-Smirnov tests indicated no significant difference ($p<0.01$) between the two distributions.

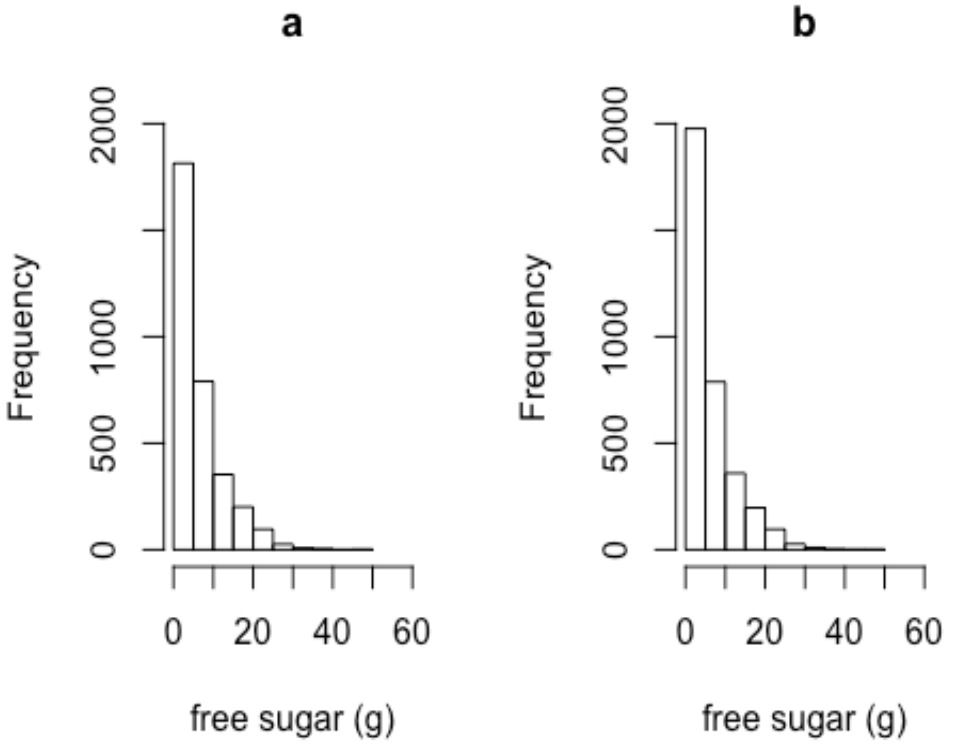


Figure 7.4a and 7.4b Comparison of distribution of free sugar estimations (g) carried out in this analysis (a) and that carried out by a previously reported mapping (b).

Table 7.1 Summary statistics for comparison of free sugar estimations (g) carried out in this analysis (a) and that carried out by a previously reported mapping (b).

	g						
	min	max	range	sum	median	mean	SE.mean
a	0	49.5	49.5	20180.6	4.3	6.1	0.1
b	0	49.5	49.5	20123.0	3.9	5.8	0.1

Table 7.2 shows the daily intake of TS and FS for 3 year olds by amount (g/d) and as percentage of TEI. The estimated mean daily intake of TS and FS were 75.8 (SD 29.3) and 40.0 (SD 23.5) g/d which contributed 26.9 (SD 5.9) and 14.1 (SD 5.81) % of TEI, respectively. The maximum daily intake reported for TS and FS was 126.0 and 77.4 g/d, respectively. Almost three-quarters of 3 year olds had FS intake greater than the WHO recommendation that FS intake is a maximum 10% of TEI while less than 4% met the lower threshold of 5% FS as % TEI.

Table 7.2 Daily intake of total and free sugar for 3 year old children by amount (g/d), frequency (as a meal or snack), as a percentage of Total Energy Intake (TEI) and the proportion of the sample population with Free Sugars (FS) intake ≥10% and ≥5% of TEI.

	Mean	SD	Minimum	Maximum
Total sugars (g/d)	75.8	29.3	34.9	126.0
Total sugar (frequency)	5.2	1.2	1.0	11.0
Energy from total sugars (%)	26.9	5.9	12.9	39.3
Free sugar (g/d)	40.0	23.5	7.2	77.4
Free sugar (frequency)	3.9	1.4	0.0	9.0
Energy from Free sugar (%)	14.1	5.81	3.1	35.3
%FS ≥ 10% (%)	74.6			
%FS ≥ 5% (%)	96.8			

The mean frequency of TS and FS consumption (as a meal or snack) was 5.2 (SD 1.2) and 3.9 (SD 1.4) times per day. The maximum frequency of consumption for TS and FS was 11.0 and 9.0 times per day, respectively. Figure 7.5 (a-d) shows the mean daily intake of all food sources of TS and FS *covered* or *non-covered* by the SFQ used in the GUI survey as g/day and as %TEI. Less than one-quarter of the mean TS intake (g/d) was *non-covered* by GUI (Figure 7.5 a) while less than one-third of the mean FS intake was *non-covered* (Figure 7.5 c). The proportions were similar when expressed as a percentage of total energy intake (Figure 7.5 b and 7.5 d).

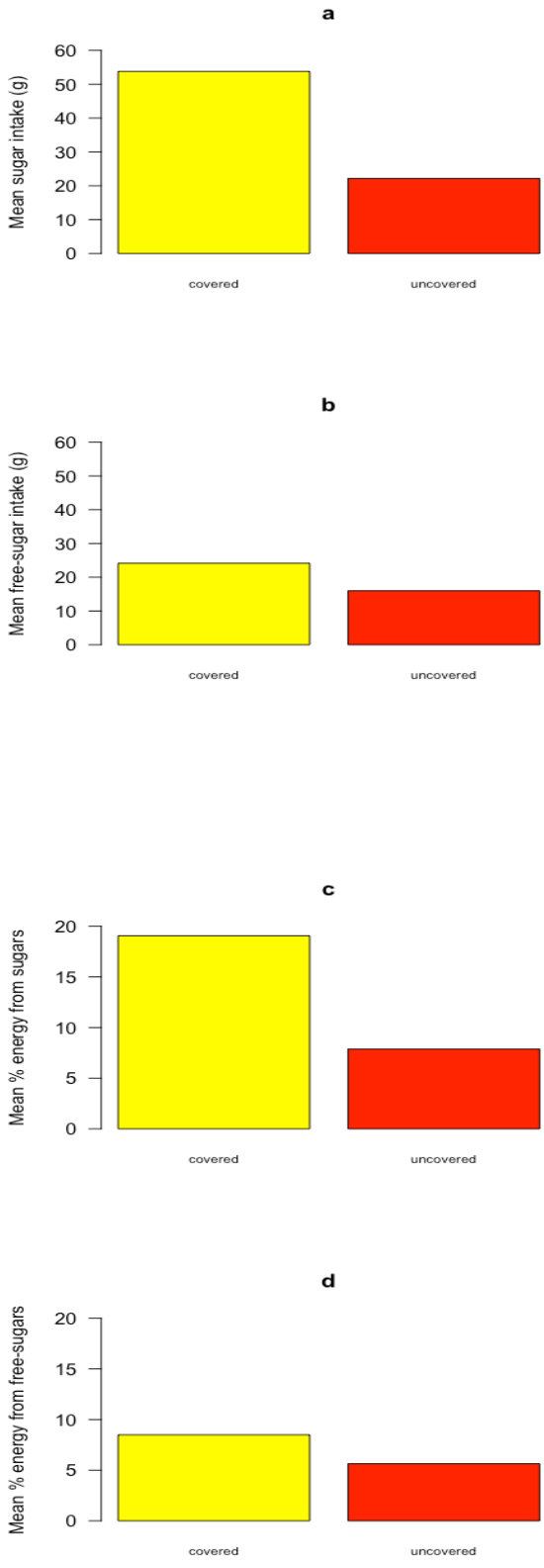


Figure 7.5 (a-d) Daily intake of total (a, c) and free sugar (b, d) *covered* and *non-covered* by GUI food groups for 3 year old children by amount (g/d) and as a percentage of Total Energy Intake (%TEI).

7.3.2. Key sugar food sources

The key sugar-contributing food sources of TS and FS intake are displayed in Table 7.3. The % consumers varied from 25% for desserts and puddings to 100% for dairy products, bread and cereals and fruit and vegetables. The most important contributors (mean g/d, %TEI), to TS intake were dairy products (22.4g/d, 7.6% TEI), fruit and vegetables (17.3g/d, 6.3and% TEI), fruit juice and smoothies (8.7g/d, 3.1% TEI) and confectionary (chocolate and non-chocolate) (5.8g/d, 2% TEI). The most important contributors to FS intake, were fruit juice and smoothies (8.4g/d, 3.0% TEI), dairy products (8.2g/d, 2.8% TEI), confectionary (chocolate and non-chocolate) (5.3g/d, 2% TEI), and soft drinks (including squashes, cordials and fruit juice drinks) (4.8g/d, 1.8% TEI). Non-chocolate confectionary and chocolate confectionary were consumed by 45% and 60% of the total NPNS sample, respectively. Dairy products were consumed by all the sample population while fruit juice and smoothies and soft drinks (including squashes, cordials and fruit juice drinks) were consumed by over 70% of the sample population. RTEBC were consumed by more 92% of children and contributed 7.8% of the total FS intake. The combination of all cakes, biscuits and confectionary contributed 37.8% of the total FS intake.

Providing dietary advice to patients regarding the reduction of CF snacks and limiting CF consumption to main mealtimes only, has long been a recommendation for the prevention of dental caries (Moynihan, 2002, Moynihan and Petersen, 2004). To give an estimate of the probability of consuming a food group as a snack or main meal all eating occasions were treated as independent occasions. For example, the count of total number of snacks as a proportion of the total number of eating occasions provided an approximate estimate of the probability of consuming a particular food group as a snack. There was a high probability (2:1) of consuming chocolate confectionary, cakes and biscuits and non-chocolate confectionary as a snack while this probability was the opposite (1:2) for consumption of fruit juices and smoothies, dairy products and soft drinks (non-diet) as a snack. RTEBC were nearly always likely to be consumed as part of a main meal.

To illustrate the effect of exclusion of all snack consumption of soft drinks (including carbonated, (non-diet), beverages, squashes, cordials and fruit juice drinks), confectionery (chocolate and non-chocolate), cakes and biscuits, sugar and syrups the analysis of % FS estimate was repeated with these items “removed” from the dietary intake would reduce the mean daily intake of FS

from 14.1% to 11%. As a population level estimate, this would double the proportion of 3 year olds meeting the maximum WHO recommendation of 10% TEI from one-quarter to one-half. However, this does not account for other changes in dietary intake that might be affected by substitution of snacking these food and drinks.

Table 7.3 Contribution of key sugar-contributing food sources to total sugar and free sugar intake in 3 year old children as weight (g/d), as a percentage of total energy intake (%TEI), by percentage consumers and by probability of consumption as part of a snack or main meal.

TEI= Total energy intake; Freq= frequency

	Fruit juices and smoothies	Dairy	Soft Drinks (Non Diet)	Chocolate confectionery	Cakes and Biscuits	Non-chocolate confectionery	Sugar and syrups	Desserts and Puddings	RTEBC	Other	Bread and Cereals	Fruit and vegetables
Consumers (%)	73.0	100.0	71.4	59.5	89.7	45.2	56.3	25.4	92.1	100.0	100.0	100.0
Total sugars.mean	8.7	22.0	4.8	3.6	4.7	2.2	2.6	1.2	3.2	3.0	2.5	17.3
Total sugars.SD	8.7	11.1	8.8	4.8	4.6	2.9	4.6	4.3	2.9	2.2	2.3	11.5
Free sugars.min	0.5	0.0	0.0	0.6	0.0	1.2	0.0	0.0	0.2	0.0	0.0	0.0
Free sugars.max	36.8	26.5	51.5	18.9	28.3	15.2	18.0	11.4	19.9	9.3	15.6	5.4
Free_sugars.mean.	8.4	8.2	4.8	3.1	4.4	2.2	2.5	0.9	3.1	1.3	0.7	0.4
Free sugars.SD	8.7	6.2	8.8	4.1	4.4	2.9	4.4	3.1	2.8	1.5	1.7	0.7
% TEI sugars.mean.	3.1	7.6	1.8	1.2	1.6	0.8	0.9	0.4	1.2	1.1	0.9	6.3
% TEI sugars.SD	3.1	3.3	3.6	1.6	1.4	1.2	1.4	1.3	1.1	0.8	0.8	4.0
%TEI free sugars.mean	3.0	2.8	1.8	1.1	1.5	0.8	0.9	0.3	1.1	0.5	0.2	0.1
% TEI free sugars.SD	3.1	1.9	3.6	1.4	1.3	1.2	1.3	0.9	1.0	0.5	0.6	0.3

Freq.mean.	0.7	3.2	0.9	0.3	0.8	0.2	0.4	0.1	0.9	5.8	1.8	3.2
Freq.SD	0.6	1.3	0.9	0.3	0.8	0.3	0.5	0.3	0.5	1.7	0.7	1.7
Prob. as Snack	27.0	26.0	30.0	73.0	69.0	66.0	18.0	37.0	5.0	21.0	19.0	33.0
Prob.as Meal	73.0	74.0	70.0	27.0	31.0	34.0	82.0	63.0	95.0	79.0	81.0	67.0

Table 7.3 (continued) Contribution of key sugar-contributing food sources to total sugar and free sugar intake in 3 year old children as weight (g/d), as a percentage of total energy intake (%TEI), by percentage consumers and by probability of consumption as part of a snack or main meal.

7.4. Discussion

In the early 1970s Ireland was one of the world's highest sugar consumers and reached a peak per capita consumption of more than 150g sugar per day (Friel et al., 1996). Currently, per capita consumption is estimated at 96.7 g/day and Ireland is still one of the highest consumers of sugar in Europe (Statista, 2016). The main aim of this paper was to estimate TS and FS intake for 3 year old children, identify the key food sources and to discuss the implications for dental health and dietary advice in the context of the recent WHO guidelines. This analysis used a modified version (Figure 2.5) of previously developed protocol (Louie et al., 2015) to map the FS content of the NPNS dataset for 3 year olds. The results (Figure 7.4a and 7.4b) suggested that while there is a degree of subjectivity in assigning FS estimates, the overall figures were consistent with a previous report using the same cohort data (Newens and Walton, 2016). The TS and FS intakes contributed to 26.9 and 14.1% of TEI, respectively. The WHO are currently reviewing recommendations for TS intake (Moynihan et al., 2018). Only a small minority (less than 4%) of children achieved the WHO conditional recommendation for the lower threshold for FS intake while three-quarters of children exceeded the higher maximum threshold of 10% TEI. Fruit juices and smoothies, dairy products (including yoghurts and fromage frais), soft drinks (including squashes, cordials and fruit juice drinks), confectionary (chocolate and non-chocolate) and cakes and biscuits were the key food sources for FS, contributing to more than three-quarters of total FS intake. For consumers only, fruit juice and smoothies, dairy products and soft drinks (including squashes, cordials and fruit juice drinks) contributed two-thirds of the total FS intake. The key sources of TS were dairy products, fruit and vegetables and fruit juice and smoothies contributing 63% of TS intake. Using these food groups and treating all snacks and main meals independently, the probability of consuming chocolate and non-chocolate confectionary and cakes and biscuits ranged from 66-73% whereas the probability of consuming fruit juice and smoothies or non-diet soft drinks (including squashes, cordials and fruit juice drinks) was 27% or 30%, respectively.

Previous results reported that older Irish children (5-12 years old) had a frequency of intake of TS of four times per day which corresponded to a mean added sugar intake of 14.6% of TEI (Joyce et al., 2008). Our results suggested that younger children (3 years old) had a mean frequency of more than 5 times per day of TS and almost 4 times per day of FS. The current advice to reduce both the amount of sugar and aim towards a maximum frequency of 1/day of sugary foods and drinks is aimed at meeting the WHO guidelines (World Health Organization, 2015, Moynihan, 2016). The SFQ in GUI did not capture approximately one-third of FS and one-quarter of TS intake suggesting that relying on modified short food questionnaires can result in significant underestimation of typical intake of food sources of sugar. This emphasises the importance of ensuring the most appropriate instrument is selected at survey design stage to achieve the optimal results within the constraints of resources. If, for example, body weight and height are the only physical parameters measured in a survey, allowing estimation of BMI and obesity, then a dietary intake instrument that can sufficiently capture total energy and habitual food intake would be appropriate. Looking at possible relationships between sugar intake and obesity without capturing one-third of FS consumption may lead to misleading conclusions.

As a prerequisite for setting guidelines, targeting public health policy and measuring adherence to recommendations it is necessary to quantify the current intake of TS/FS/AS and the main food sources (Lei et al., 2016, Sluik et al., 2016, Ruiz et al., 2017, Gibson et al., 2016, Erickson and Slavin, 2015b, Azais-Braesco et al., 2017, Newens and Walton, 2016). Although there are very few data on FS levels of consumption, especially at this age, our results were similar to those previously reported where a large proportion of children greatly exceeded the recommended 10% FS as energy intake (Lei et al., 2016, Sluik et al., 2016, Ruiz et al., 2017, Azais-Braesco et al., 2017, Gibson et al., 2016, World Health Organization, 2017a). Although most of the data available in the EU has reported on AS intake, rather than FS, this suggests, given the more narrowly defined AS (Figure 1.4), that the level of FS intake is probably even higher. For example, AS intake in the USA in 2-5-year olds is 13.4% of TEI,

however, as noted in a recent commentary (Moynihan et al., 2018), these figures exclude FS in 100% natural fruit juice. In this analysis, the mean free sugar intake was 40g/day, similar to that reported for AS intake for 4-year old children in the UK and Denmark (World Health Organization, 2017a). Furthermore, as dietary surveys tend to under-estimate sugar intake, FS consumption is, probably, under-reported (Livingstone and Black, 2003).

Analytically, it is not possible to distinguish TS and FS (Erickson and Slavin, 2015b, Stephen et al., 2012, Louie et al., 2015) and most methods to estimate FS have a degree of subjectivity and variation between country and product (Sluik et al., 2016, Ruiz et al., 2017, Gibson et al., 2016). Currently, food manufacturers are not required to include FS content in their labelling. The problem is compounded by the lack of standardisation of terminology and methods adds to the well-known difficulties that already exist in measuring food intake in young children (Magarey et al., 2011). While the leading sources of FS or AS intake tend to be the low nutrient, discretionary foods (Lei et al., 2016, Welsh and Figueroa, 2017, Sluik et al., 2016, Azais-Braesco et al., 2017) our analysis suggested that some nutrient-rich foods, such as sweetened yoghurts (dairy products), are also significant contributors of FS at this young age. Furthermore, even within the EU there are large variations between countries in the sugar content of some of these energy-dense, discretionary foods (World Health Organization, 2017a). For example, RTEBC has been highlighted as a key source of AS/FS intake with efforts being made to reformulate these products as part of reducing overall FS consumption (Public Health England, 2017, Public Health England, 2015b, World Health Organization, 2017a). However, there are wide variations within and across countries in the sugar content of RTEBC and in our analysis these products contributed less than 8% of total FS intake. Thus, to reduce FS intake it is important to consider the most efficient products for reformulation given variations in content between countries and in consumption patterns at different ages (Public Health England, 2017, Public Health England, 2015b, World Health Organization, 2017a).

Most analyses of cross-sectional data have reported an inverse association between the overall level of FS/AS intake and nutrient density (Joyce et al.,

2008, Gibson et al., 2016). Recent evidence from a prospective study of young children in Australia indicated that carbonated soft drinks (non-diet) may have increased cariogenic potential due to higher acidity and added buffering agents that can prolong a lower pH environment (Gussy et al., 2016). This is a key age for changes in discretionary food intake (Johnson et al., 2016) and recent studies (Chaffee et al., 2015, Amezdroz et al., 2015) have demonstrated the association between increased exposure to cariogenic food items in infancy and dental caries in the later preschool age. A number of researchers have suggested that food-based guidelines should focus on discouraging energy dense, nutrient-poor, discretionary food sources such as confectionary and soft drinks as a more practical approach to gaining a better overall nutrient intake than restricting all foods containing sugar (Gibson et al., 2016, Erickson and Slavin, 2015b, Lei et al., 2016). This seems to be a sensible approach, particularly for age ranges where other key food sources such as RTEBC and sweetened dairy products still contribute significantly to overall nutrient intake. Exclusion of all snack consumption of soft drinks (including carbonated, (non-diet), beverages, squashes, cordials and fruit juice drinks), confectionery (chocolate and non-chocolate), cakes and biscuits, sugar and syrups from the diet of the 3 year olds in this analysis would reduce the mean daily intake of FS from 14.1% to 11%. As a population level estimate, this would double the proportion of 3 year olds, from one-quarter to one-half, meeting the maximum WHO recommendation of 10% TEI. However, this does not account for other changes in dietary intake that might be affected by substitution of snacking these food and drinks.

Given the high levels of FS intake relative to the WHO recommendations, even at this young age, it would appear to be an opportunity for all health care professionals to focus on strategies to affect food intake behaviour. Unfortunately, there is a deficit of research on the measurement and effectiveness of dietary interventions in the dental practice (Arheiam et al., 2016a, Arheiam et al., 2016b, Meyer and Lee, 2015). As there is also a lack of high-quality cohort studies that measure dental caries levels and the frequency and amount of FS intake (Moynihan et al., 2018, Moynihan, 2016) it is,

currently, difficult to advance the state of knowledge in this crucial aspect of primary disease prevention. While reducing the frequency of consumption of FS can assist in lowering dental caries risk it is also necessary to reduce the amount of FS to reduce the risk of other non-communicable diseases related to excess sugar intake (Moynihan, 2016, World Health Organization, 2015).

Evidence supporting the updated WHO guidelines were primarily based on a systematic review of the relationship between sugar intake and dental caries (Moynihan and Kelly, 2014) and that between sugar and obesity (Te Morenga et al., 2013). The authors noted that limiting FS intake to, ideally, less than 5% TEI, at a population level reduces the risk of dental caries throughout the life-course. As dental caries is a cumulative, dynamic disease process even a small reduction in risk in early childhood can have significance later in life (Gussy et al., 2016, Moynihan and Kelly, 2014). However, although WHO have made recommendations for the amount (World Health Organization, 2015) and frequency (World Health Organization, 2003) of sugar intake there are many other factors that affect the cariogenicity of food and drink for children such as timing, infant feeding practices, duration in the oral cavity, saliva, fluoride exposure and the composition of the plaque biofilm (Bradshaw and Lynch, 2013, Chaffee et al., 2015, Burt et al., 1988, Johansson et al., 2010, Moynihan and Petersen, 2004). Interestingly, a recent systematic review (Erickson et al., 2017) has questioned the scientific basis and low strength of evidence for recent guidelines on sugar intake (World Health Organization, 2015) and emphasised the importance of understanding these limitations when dietary behaviour is being considered.

Apart from the limitations of measuring dietary intake already outlined in Section 5 there are other issues with terminology when researchers are trying to compare the frequency of intake of TS/FS including the definition of eating occasions or snacks (Leech et al., 2015, Marshall et al., 2005). The lack of standardisation of commonly used terms plus the known phenomenon of under-reporting of snacks, particularly by subjects who are obese or overweight, will influence the number of EO recorded and reported (Leech et al., 2015) (Gibney and wolever, 1997). The limitations of methods to assess

dietary intake (Thompson and Subar, 2013), particularly in very young children (Magarey et al., 2011), are outlined in Section 1.5.2. Individuals also tend to reduce their actual consumption when intake is monitored, although for parent-reported data this may not be as problematic. However, parental recall of food intake is likely to lead to under-reporting (Magarey et al., 2011) and this may be more pronounced with foods considered to be unhealthy. Under-reporters were not excluded from this analysis. As suggested in a recent review of TS and AS intakes in Europe (Azais-Braesco et al., 2017) there is an urgent need to develop a standardised systematic methodology, similar to that developed by Louie et al (2015), to minimise the reporting of inappropriate estimates of AS or FS. However, our analysis was based on nationally representative data and used food intake estimates measured using a 4 day weighed food diary. FS estimates were similar to those previously reported for this age group.

7.5. Conclusions

Accurate and reliable data on FS intake at the preschool age is a limiting factor in understanding consumption levels and pattern of food sources. This analysis highlights the usefulness in adopting a consistent approach to FS estimation and the importance of using appropriate methods for determining sugar intake at the food level. Brief SFQ's are not suitable techniques for understanding habitual sugar intake and will lead to bias when examining possible associations with disease. A large majority of 3 year old Irish children do not meet the WHO recommended guidelines for FS intake and almost none meet the desired conditional recommendation. Consequently, it can be reasoned that FS intake is excessively high even at this early age and targeting low nutrient discretionary food and drink seems a reasonable approach to achieving an overall reduction in FS consumption.

Chapter 8. General Discussion

Despite improved levels of oral health in the general population over previous decades, the prevalence of oral health problems in younger children has increased in recent years (Bourgeois and Llodra, 2014, Dye et al., 2010). The primary caregiver (PCG) is the gate-keeper in providing general and oral health care for the developing child. The last Child Oral Health Survey in Ireland, which did not include preschoolers, reported that approximately 'one-in-three' 5-year-old children had a dental caries experience; and more than 80% of the caries experience was 'untreated' (Whelton et al., 2006). This poses a significant health burden to children and can result in pain, infection, abscess formation and require restorative treatment or surgery (frequently under general anaesthesia) and also affects basic functioning – physically, socially and psychologically (Gussy et al., 2016, Gussy et al., 2006, Boeira et al., 2012, Finlayson et al., 2007).

There are a number of distinct issues in attempting to research preschoolers oral and general health predictors including reliance on the PCG for reporting information, a general lack of availability of health-related data for this age group (Bonotto et al., 2017, Daher et al., 2015, Lo et al., 2014) and specific difficulties in accurately measuring dietary intake (Magarey et al., 2011, Thompson and Subar, 2013). Recently, the WHO have recommended the inclusion of the 3 year old age group as one of the index ages for population surveys in the next edition of the WHO's *Oral health surveys: basic methods* (World Health Organization, 2017b) reflecting the new understanding of how important this preschool period is in establishing oral-health related behaviour and dietary preferences that impact on future health and development (Amezdroz et al., 2015, Johnson et al., 2016, Gussy et al., 2016). Advances in omics technology and biological, psychosocial, behavioural, and data science have created a wealth of research knowledge that has yet to be fully realised in attaining the highest standards of health in young children. There has been recognition of the need for a more integrated approach to dealing with one of

the most common and costly health problems in childhood- oral disease (Divaris, 2016, Casamassimo et al., 2014) and how the latest findings can be translated into improved clinical practice and population health (Sniehotta et al., 2017, Meyer and Lee, 2015).

This research explored multiple variables based on their relevance to child dental health, used the conceptual framework of Fisher-Owens (2007) and applied innovative new approaches utilising the powerful tools of data science (Binder and Blettner, 2015, Divaris, 2016, Krebs-Smith et al., 2015). The data sources selected for our secondary analyses were from two nationally representative preschool cohorts, the GUI and NPNS. GUI was selected as it is Ireland's most substantive research initiative ever undertaken of childrens' health and development and includes multiple psychosocial, behavioural and environmental factors that impact on oral health outcomes. NPNS was used, when, as the project progressed, it was identified that a more detailed food database was required to understand cariogenic food intake of 3 year olds.

Exploratory data analysis (EDA) was carried out using both traditional statistical approaches e.g., logistic regression (Agresti, 2007), and modern data science tools, e.g., supervised learning methods (Leskovec et al., 2014, Maimon and Rokach, 2009). Recently, there has been a convergence between EDA and data mining and this has been accelerated by the use of more powerful visualisation tools and a change in the more traditional data modelling culture (Ho Yu, 2010). Using this data science approach and non-parametric methods including decision trees, this study initially focussed on risk indicators and dental problems in children. Essentially, EDA was used in the new context of data science which relies strongly on data visualisation and predictive modelling (Peng, 2012, Peng and Matsui, 2015, Wickham and Grolemund, 2016, Ho Yu, 2010, Maimon and Rokach, 2009).

Recent commentaries (Lee and Divaris, 2014, Sniehotta et al., 2017, Divaris, 2016) highlight the need for an integrated multidisciplinary approach to population based health/oral health research rather than individually examining the effects of upstream or downstream risk factors (Krieger, 2008). This analysis has illustrated both the usefulness of using an interdisciplinary

database and the effectiveness of using CTA and other data visualisation techniques to highlight multilevel interactions. In this study, the resulting tree model outputs were relatively easy to interpret, useful for identifying important variables and demonstrated that PCG and child psychosocial and general health factors were associated with early childhood dental problems, even before the completion of the primary dentition. Given the lack of knowledge of food consumption patterns at this young age, the second key aim, using a common risk factor approach, was to investigate weight status and dietary intake of CF. The study of CF intake was further explored by developing novel data mapping techniques and using other data mining methods from the retail industry, such as rule association analysis (Höppner, 2005). It is remarkable, given the importance of CF, and FS specifically, in contributing to early childhood caries that greater emphasis has not been placed on understanding these meal/snack patterns of consumption and interactions between food components.

While overweight and obesity dominate the focus of recent research with children, the results of this analysis suggested that it is also important to consider underweight (thinness) in early childhood as a condition related to poor health outcomes. Adopting a common risk factor approach to diet, weight status and dental caries was proposed almost two decades ago (Sheiham and Watt, 2000) but the relationship is still unclear (Hooley et al., 2012a). The prevalence of underweight and obesity was similar in the 3 year old GUI cohort. The results of CTA of almost 9,800 children indicated that a subgroup of children, which combined obese and underweight categories, had the highest prevalence of dental problems (19%). However, this was a small subgroup and only illustrated the difficulties in investigating weight status, dietary intake and dental problems, especially at this young age, using a cross-sectional approach. Ethnicity was the most important predictor of dental problems in this CTA, which was particularly interesting as this has been used as a variable in relatively few studies of dental caries in children. The PCG education level, although not a significant predictor in this study's CTA model, has been consistently shown to be an important risk factor for caries in children (Harris

et al., 2004). Ultimately, the common risk factor approach may be a pragmatic means of developing shared modifiable strategies for prevention of both dental and weight problems. To progress our knowledge in these areas and develop evidence-based policy approaches to oral health care it is crucial to have large high-quality datasets with both medical and dental information that can be readily analysed (Carson et al., 2017). However, a degree of caution needs to be applied to the application and interpretation of predictive models and adopting a 'black-box' approach is inadvisable (Breiman, 2001, Ho Yu, 2010). Data mining algorithms generally favour high overall classification without any regard for individual class significance. This has been referred to as the accuracy paradox, where the accuracy measure is high but only reflects the underlying class distribution (Sun et al., 2009). This underlines the importance of reporting more performance indicators than predictive accuracy alone and being aware of the statistical properties of the data distribution. Thus, using new analytic software or techniques without a clear understanding of the statistical logic behind the methods can result in inappropriate analysis and questionable results (Hastie et al., 2009, Ho Yu, 2010).

The results of initial analyses prompted the need for more detailed food intake data for 3 year old children and highlighted the importance of selecting the most appropriate dietary instrument at survey design stage. To augment the limited dietary intake data in GUI the detailed NPNS data was used to map the food categories in GUI. This protocol provided a method for further mapping of national cohort surveys and food databases for other age cohorts. Through mapping the food codes and estimating the degree of *non-covered* food it was possible to visualise the relative performance of the Short Frequency Questionnaire (SFQ) compared to the more detailed one especially in capturing specific food types, e.g., high sugar foods. The SFQ did not capture a substantial portion of habitual food intake of 3 year olds in Ireland. Researchers interested in focussing on specific foods, such as those high in sugar, could use this approach in the future to easily assess the proportion of foods *covered* or *non-covered* by reference to the mapped, more detailed, food database.

The consumption of CF was estimated, for consumers only, using a mean daily intake method and an average consumption method. Differences in distribution patterns of snack and main meal consumptions were easily visualised using asymmetric bean plots. This illustrated how the method used to estimate and report food intake can influence the interpretation of results. Association analysis, was used to identify the food components of all eating occasions and the combinations of CF and non-CF (NCF) components in the most commonly consumed meals and snacks. Using these descriptors (cariogenic and non-cariogenic aggregates), the most commonly consumed CF items as snacks were also low nutrient discretionary foods and drinks that were key contributors to FS intake. Thus, efforts to reduce FS intake should first target these discretionary items that contribute little to overall nutrient intake. Using similar data from the association analysis, alluvial plots visualised the CF and NCF component interactions of the 10 most frequent eating occasions of cariogenic food. This provided an insight into how CF components are consumed with (or without) NCF components in meals and snacks which can provide useful information on dietary advice for dental health. Each meal was composed of 3.4 food components and one-quarter of these components were CF. In general, our results suggested that the overall intake of CF was high compared to recommended dietary guidelines for children.

A recent commentary noted that nutrition research is one of the most contentious areas in science (Ioannidis and Trepanowski, 2018). There are a number of limitations when investigating the effect of individual foods or nutrients on oral health, not least of which is the capture of accurate dietary intake data. As mentioned previously, brief SFQs are commonly used in large surveys such as that used in GUI. This study has clearly demonstrated the unsuitability of these instruments for understanding habitual food intake and the potential bias when examining possible associations. Although there is no accepted valid approach for FS estimation, the protocol used to map FS intake in this research found similar results to previously reported results (Walton et al., 2016). A large majority of 3 year old Irish children did not meet the WHO recommended guidelines for free sugar intake and almost none met the desired

conditional recommendation of 5% total energy intake from FS. Consequently, it can be reasoned that free sugar intake is excessively high even at this early age and targeting low nutrient, discretionary food and drink seems a reasonable approach to achieving an overall reduction in consumption. For example, using the NPNS data, if all snack consumption of soft drinks (including carbonated, (non-diet), beverages, squashes, cordials and fruit juice drinks), confectionery (chocolate and non-chocolate), cakes and biscuits, sugar and syrups were excluded from the diet this would reduce the mean daily intake of FS would be reduced from 14.1 to 11% for the total sample population. As a population level estimate, this would double the proportion of 3 year olds meeting the maximum WHO recommendation of 10% TEI from one-quarter to one-half. However, this does not account for other changes in dietary intake that might be affected by substitution of snacking on these foods and drinks. Given that eating habits developed at this age influence eating behaviours over the lifecourse (Gibson et al., 2012), this may, however, offer a potential focus for public health services in Ireland to further implement a common risk factor approach to obesity and dental disease.

Methodological difficulties in conducting both dietary assessment (Ioannidis, 2013, Thompson and Subar, 2013) and oral health research (O'Mullane et al., 2012) are considerable. Developing techniques such as data mapping and data linkage protocols can help maximise the potential within datasets and foster interdisciplinary approaches (Slack-Smith, 2012, Schenker and Raghunathan, 2007). The NPNS data is the only detailed food database for preschool children in Ireland and, to the best of our knowledge, this analysis was the first to use data mapping protocols for the purpose of exploiting the value of a larger dataset (GUI) of children of a similar age. However, the future of nutritional epidemiology will revolve around “real-time” methods that capture habitual and changing dietary intake over time (Krebs-Smith et al., 2015, Boushey et al., 2017, Forster et al., 2014). This will also require continued development of standardised terminology and application of appropriate data-driven or investigator defined techniques (Krebs-Smith et al., 2015) for measuring dietary

patterns, even including the seemingly mundane definition of the humble 'snack' (Leech et al., 2015).

Applying "big data" techniques to evaluate observational data can be regarded as a form of data mining which is effectively the discovery of "models" for data (Leskovec et al., 2014). Training in the application of these techniques is essential if dental epidemiology is to equip future researchers with the skills needed to meet the requirements of modern epidemiological research. Integration of "big data" into the practice of epidemiology was one of the key recommendations to transform epidemiology for 21st Century Public Health Medicine (Khoury et al., 2013). This is already occurring in areas such as bioinformatics and genomics (Lee et al., 2008, Binder and Blettner, 2015). Constant argument over upstream and downstream factors to decide upon and which public health approach to support, i.e., population-level versus individual-level, has not advanced the science of public health (Baker and Gibson, 2014). As suggested recently (Sniehotta et al., 2017), a new model is required for public health that is built on a comprehensive and sophisticated knowledge of complex systems (Luke and Stamatakis, 2012). However, successful targeting of high risk subgroups would require a risk prediction model with both high sensitivity and specificity. Machine learning algorithms can improve predictive accuracy over more conventional regression models and, as demonstrated here, can illustrate nonlinear interactions between variables (Chen and Asch, 2017). It is essential to be cautious when using these powerful analytical techniques and have a solid understanding of the underlying domain knowledge and the data characteristics/ distribution to avoid overfitting the data. High-quality longitudinal cohort data and valid methodology must be the cornerstone for creating accurate risk assessment models (Divaris et al., 2017, Chen and Asch, 2017).

Data analysis involves the transformation of raw data into usable information (Binder et al., 2006). Data visualisation is an exciting branch of data science (Wegman, 2003) that appears to have been underutilised in dental research to date and these techniques, which can be complex to apply and comprehend initially, hold large potential for identifying useful patterns and associations in

large datasets (Lander, 2014, Williams, 2011, Tufte, 2001). As demonstrated in this research, the use of bean plots or alluvial diagrams to visualise data can highlight patterns within the data that are difficult to otherwise identify. This was elegantly emphasised by Tufte's (2001) classic portrayal of Anscombe's quartet which showed that graphics can be more revealing than conventional statistical parameters where all four datasets described by the same linear regression model have very different patterns of distributions.

It is particularly important to develop code writing skills so that the data analysis process is logical and transparent and keeps a clear and reproducible record. As demonstrated in this analysis, R Markdown was very useful for creating code documents and provided a clear record of the text and code generated. Any subsequent alterations to the input data or code chunks can be re-executed and automatically regenerate documents in Word®, HTML or PDF formats (Lander, 2014).

Inevitably, there were difficulties with this project. Full access to the researcher microdata files from GUI required appointment as an Officer of Statistics. While the practical and conceptual impetus for interdisciplinary research is well recognised (Van Noorden, 2015) there are a number of resistances encountered when submitting papers, applying for grants and collaborating (MacLeod, 2018). Linking distinct datasets at a national level is desirable from a research perspective (Kim and Rao, 2011, Schenker and Raghunathan, 2007). Public Health England have proposed to link (using anonymised ID) the National Diet and Nutrition Survey with the National Oral Health Survey for the purpose of investigating the common risk factor approach to overweight/obesity and oral health (Public Health England, 2013b). However, all aspects of data usage will need to be reviewed under the new general data protection regulations that come into effect in May 2018.

In conclusion, the research completed in this thesis presents a detailed set of analyses of key risk indicators of dental health problems in preschool children with an emphasis on using non-traditional approaches adopted from data science. Beginning with classification tree models, the association between the prevalence of dental problems in preschoolers and health and psychosocial

characteristics of the PCG and child were demonstrated. Subsequently, through focussing on the common risk factor approach to weight status and dental health, the later Sections of the research concentrated on developing novel methods for exploring the consumption of cariogenic food in 3 year olds in Ireland. The code developed for the mapping protocols for the food database and for FS estimation has provided an opportunity for other researchers to utilise these techniques or copy code fragments that may be applied to other food databases while some of the visualisation techniques gave an insight into how beneficial these tools are for highlighting trends that may be otherwise hidden within the data.

The findings presented highlighted the relatively high prevalence of reported dental problem visits in early childhood, with 1 in 50 children at 9 months and 1 in 20 at 3 years of age having a reported dental problem that resulted in the PCG seeking child health care. However, to advance the usefulness of this data, further proposals have been submitted for a nested study to the Department of Children and Youth Affairs (DCYA). The GUI infant cohort includes specific information on oral health, however, clinical data on dental caries experience is lacking. The ability to obtain clinical information on a group of children 'nested' in this cohort offers a unique opportunity to examine 'how' and 'why' dental caries develops. Given the anticipated prevalence rates of dental caries and the model analysis already completed on the existing data, it is intended to use importance sampling to obtain a 'nested' group within the existing infant cohort who would now be, approximately, 10 years old. To pursue the theme of the current research multiple oral health survey questions have been successfully incorporated into the next National Child Food Survey which will provide more detailed nutritional data that can be directly explored in relation to dental variables for children 5-12 years of age. Building on previous analyses this data could be explored using new R software packages such as caret (classification and regression training) that attempt to streamline the process for creating predictive models. The future of dental epidemiology must be closely allied with other disciplines, particularly those in the data,

behavioural and nutritional sciences, which can focus on modifiable risk changes.

References

ABREU, L. G., ELYASI, M., BADRI, P., PAIVA, S. M., FLORES-MIR, C. & AMIN, M. 2015. Factors associated with the development of dental caries in children and adolescents in studies employing the life course approach: a systematic review. *Eur J Oral Sci*.

AGRESTI, A. 2007. Logistic regression. *An Introduction to Categorical Data Analysis, Second Edition.* New York: Wiley, 99-136.

AHRENS, W., PIGEOT, I., POHLABELN, H., DE HENAUW, S., LISSNER, L., MOLNAR, D., MORENO, L. A., TORNARITIS, M., VEIDEBAUM, T., SIANI, A. & CONSORTIUM, I. 2014. Prevalence of overweight and obesity in European children below the age of 10. *Int J Obes (Lond),* 38 Suppl 2**,** S99-107.

ALLISON, D. B., BASSAGANYA-RIERA, J., BURLINGAME, B., BROWN, A. W., LE COUTRE, J., DICKSON, S. L., VAN EDEN, W., GARSSEN, J., HONTECILLAS, R., KHOO, C. S., KNORR, D., KUSSMANN, M., MAGISTRETTI, P. J., MEHTA, T., MEULE, A., RYCHLIK, M. & VOGELE, C. 2015. Goals in Nutrition Science 2015-2020. *Front Nutr,* 2**,** 26.

ALVES, L. S., SUSIN, C., DAME-TEIXEIRA, N. & MALTZ, M. 2013. Overweight and obesity are not associated with dental caries among 12-year-old South Brazilian schoolchildren. *Community Dent Oral Epidemiol,* 41**,** 224-31.

AMERICAN ACADEMY OF PEDIATRIC DENTISTRY 2013. Guideline on periodicity of examination, preventive dental services, anticipatory guidance/counseling, and oral treatment for infants, children, and adolescents. *Pediatr Dent,* 35**,** E148.

AMERICAN ACADEMY OF PEDIATRIC DENTISTRY 2016. Policy on Early Childhood Caries (ECC): Classifications, Consequences, and Preventive Strategies. *Pediatr Dent,* 38**,** 52-54.

AMEZDROZ, E., CARPENTER, L., O'CALLAGHAN, E., JOHNSON, S. & WATERS, E. 2015. Transition from milks to the introduction of solid foods across the first 2 years of life: findings from an Australian birth cohort study. *Journal of human nutrition and dietetics,* 28**,** 375-383.

AMINABADI, N. A., GHOREISHIZADEH, A., GHOREISHIZADEH, M., OSKOUEI, S. G. & GHOJAZADEH, M. 2014. Can Child Temperament Be Related to Early Childhood Caries? *Caries research,* 48**,** 3-12.

AMINE, E., BABA, N., BELHADJ, M., DEURENBERY-YAP, M., DJAZAYERY, A., FORRESTER, T., GALUSKA, D., HERMAN, S., JAMES, W. & MBUYAMBA, J. 2002. *Diet, nutrition and the prevention of chronic*

diseases: report of a Joint WHO/FAO Expert Consultation, World Health Organization.

ANDERSON, A. S. 2014. Sugars and health - risk assessment to risk management. *Public Health Nutr,* 17**,** 2148-50.

ANDERSON, C., CURZON, M., VAN LOVEREN, C., TATSI, C. & DUGGAL, M. 2009. Sucrose and dental caries: a review of the evidence. *Obesity reviews,* 10**,** 41-54.

ANDERSON, S. E., RAMSDEN, M. & KAYE, G. 2016. Diet qualities: healthy and unhealthy aspects of diet quality in preschool children. *Am J Clin Nutr,* 103**,** 1507-13.

ARCELLA, D., OTTOLENGHI, L., POLIMENI, A. & LECLERCQ, C. 2002. The relationship between frequency of carbohydrates intake and dental caries: a cross-sectional study in Italian teenagers. *Public Health Nutr,* 5**,** 553-60.

ARCHER, E. & BLAIR, S. N. 2015. Implausible data, false memories, and the status quo in dietary assessment. *Adv Nutr,* 6**,** 229-30.

ARHEIAM, A., ALBADRI, S., BROWN, S., BURNSIDE, G., HIGHAM, S. & HARRIS, R. 2016a. Are diet diaries of value in recording dietary intake of sugars? A retrospective analysis of completion rates and information quality. *Br Dent J,* 221**,** 571-576.

ARHEIAM, A., BROWN, S. L., HIGHAM, S. M., ALBADRI, S. & HARRIS, R. V. 2016b. The information filter: how dentists use diet diary information to give patients clear and simple advice. *Community Dent Oral Epidemiol,* 44**,** 592-601.

AZAIS-BRAESCO, V., SLUIK, D., MAILLOT, M., KOK, F. & MORENO, L. A. 2017. A review of total & added sugar intakes and dietary sources in Europe. *Nutr J,* 16**,** 6.

BAGGIO, S., ABARCA, M., BODENMANN, P., GEHRI, M. & MADRID, C. 2015. Early childhood caries in Switzerland: a marker of social inequalities. *BMC Oral Health,* 15**,** 82.

BAGRAMIAN, R. A., GARCIA-GODOY, F. & VOLPE, A. R. 2009. The global increase in dental caries. A pending public health crisis. *Am J Dent,* 22**,** 3-8.

BAKER, S. R. & GIBSON, B. G. 2014. Social oral epidemi(olog)(2) y where next: one small step or one giant leap? *Community Dent Oral Epidemiol,* 42**,** 481-94.

BARBOUR, M. 2009. Formulating tooth-friendly beverages, confectionery and oral care products. *In:* WILSON, M. (ed.) *Food Constituents and Oral Health.* Oxford: Elsevier, 470-487.

BARCLAY, A. W. & BRAND-MILLER, J. 2011. The Australian paradox: a substantial decline in sugars intake over the same timeframe that overweight and obesity have increased. *Nutrients,* 3**,** 491-504.

BATES, J. E., FREELAND, C. A. B. & LOUNSBURY, M. L. 1979. Measurement of infant difficultness. *Child development***,** 794-803.

BECK, J. D. 1998. Risk revisited. *Community Dentistry and Oral Epidemiology,* 26**,** 220-225.

BECKER, W. & HELSING, E. 1991. *Food and health data: their use in nutrition policy-making*. WHO Regional Publication, European Series. WHO Regional Office for Europe.

BEN-SHLOMO, Y. & KUH, D. 2002. A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. *Int J Epidemiol,* 31**,** 285-293.

BERNABE, E., VEHKALAHTI, M. M., SHEIHAM, A., LUNDQVIST, A. & SUOMINEN, A. L. 2015. The Shape of the Dose-Response Relationship between Sugars and Caries in Adults. *Journal of dental research,* 95**,** 167-172.

BERRY, J. O. & JONES, W. H. 1995. The parental stress scale: Initial psychometric evidence. *Journal of Social and Personal Relationships,* 12**,** 463-472.

BHUTTA, Z. A., DAS, J. K., RIZVI, A., GAFFEY, M. F., WALKER, N., HORTON, S., WEBB, P., LARTEY, A. & BLACK, R. E. 2013. Evidence-based interventions for improvement of maternal and child nutrition: what can be done and at what cost? *The Lancet,* 382**,** 452-477.

BINDER, D., ROBERTS, G. & CANADA, S. 2006. Approaches for analyzing survey data: a discussion. *ASA Proceedings of Survey Research Methods Section***,** 2711-2778.

BINDER, H. & BLETTNER, M. 2015. Big data in medical science--a biostatistical view. *Dtsch Arztebl Int,* 112**,** 137-42.

BIRO, G., HULSHOF, K., OVESEN, L. & CRUZ, J. A. 2002. Selection of methodology to assess food intake. *Eur J Clin Nutr,* 56**,** S25.

BLUFORD, D. A., SHERRY, B. & SCANLON, K. S. 2007. Interventions to prevent or treat obesity in preschool children: a review of evaluated programs. *Obesity,* 15**,** 1356-1372.

BOEIRA, G., CORREA, M. B., PERES, K., PERES, M., SANTOS, I. S., MATIJASEVICH, A., BARROS, A. J. & DEMARCO, F. F. 2012. Caries is the main cause for dental pain in childhood: findings from a birth cohort. *Caries research,* 46**,** 488-495.

BÖNECKER, M., ABANTO, J., TELLO, G. & OLIVEIRA, L. B. 2012. Impact of dental caries on preschool children's quality of life: an update. *Braz Oral Res,* 26**,** 103-107.

BONOTTO, D. V., MONTES, G. R., FERREIRA, F. M., ASSUNCAO, L. R. & FRAIZ, F. C. 2017. Association of parental attitudes at mealtime and snack limits with the prevalence of untreated dental caries among preschool children. *Appetite,* 108**,** 450-455.

BOURGEOIS, D. M. & LLODRA, J. C. 2014. Global burden of dental condition among children in nine countries participating in an international oral health promotion programme, 2012–2013. *International Dental Journal,* 64**,** 27-34.

BOUSHEY, C. J., SPODEN, M., ZHU, F. M., DELP, E. J. & KERR, D. A. 2017. New mobile methods for dietary assessment: review of image-assisted and image-based dietary assessment methods. *Proc Nutr Soc,* 76**,** 283-294.

BOWMAN, A. W. & AZZALINI, A. 1997. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, OUP Oxford.

BOX, G. E. 1976. Science and statistics. *Journal of the American Statistical Association,* 71**,** 791-799.

BOYCE, W. T., DEN BESTEN, P. K., STAMPERDAHL, J., ZHAN, L., JIANG, Y., ADLER, N. E. & FEATHERSTONE, J. D. 2010. Social inequalities in childhood dental caries: the convergent roles of stress, bacteria and disadvantage. *Social Science & Medicine,* 71**,** 1644-1652.

BRADSHAW, D. J. & LYNCH, R. J. M. 2013. Diet and the microbial aetiology of dental caries: new paradigms. *International Dental Journal,* 63**,** 64-72.

BRAND-MILLER, J. C. & BARCLAY, A. W. 2017. Declining consumption of added sugars and sugar-sweetened beverages in Australia: a challenge for obesity prevention. *Am J Clin Nutr,* 105**,** 854-863.

BRAY, G. A. & POPKIN, B. M. 2014. Dietary sugar and body weight: have we reached a crisis in the epidemic of obesity and diabetes?: health be damned! Pour on the sugar. *Diabetes Care,* 37**,** 950-6.

BREIMAN, L. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science,* 16**,** 199-231.

BREIMAN, L., FRIEDMAN, J., STONE, C. J. & OLSHEN, R. A. 1984. *Classification and regression trees,* Florida, CRC press.

BRIGHT, M. A., ALFORD, S. M., HINOJOSA, M. S., KNAPP, C. & FERNANDEZ-BACA, D. E. 2015. Adverse childhood experiences and dental health in children and adolescents. *Community Dent Oral Epidemiol,* 43**,** 193-9.

BRISBOIS, T. D., MARSDEN, S. L., ANDERSON, G. H. & SIEVENPIPER, J. L. 2014. Estimated intakes and sources of total and added sugars in the Canadian diet. *Nutrients,* 6**,** 1899-912.

BRONFENBRENNER, U. & MORRIS, P. A. 2007. The Bioecological Model of Human Development. *Handbook of Child Psychology.* John Wiley & Sons, Inc.

BURT, B. A. 2001. Definitions of risk. *J Dent Educ,* 65**,** 1007-1008.

BURT, B. A. 2005. Concepts of risk in dental public health. *Community Dentistry and Oral Epidemiology,* 33**,** 240-247.

BURT, B. A., EKLUND, S. A., MORGAN, K. J., LARKIN, F. E., GUIRE, K. E., BROWN, L. O. & WEINTRAUB, J. A. 1988. The effects of sugars intake and frequency of ingestion on dental caries increment in a three-year longitudinal study. *J Dent Res,* 67**,** 1422-9.

CAMP, N. J. & SLATTERY, M. L. 2002. Classification tree analysis: a statistical tool to investigate risk factor interactions with an example for colon cancer (United States). *Cancer Causes & Control,* 13**,** 813-823.

CANTORAL, A., TELLEZ-ROJO, M. M., ETTINGER, A. S., HU, H., HERNANDEZ-AVILA, M. & PETERSON, K. 2016. Early introduction and cumulative consumption of sugar-sweetened beverages during the pre-school period and risk of obesity at 8-14 years of age. *Pediatr Obes,* 11**,** 68-74.

CARROLL, R. J., MIDTHUNE, D., SUBAR, A. F., SHUMAKOVICH, M., FREEDMAN, L. S., THOMPSON, F. E. & KIPNIS, V. 2012. Taking advantage of the strengths of 2 different dietary assessment instruments to improve intake estimates for nutritional epidemiology. *Am J Epidemiol,* 175**,** 340-7.

CARSON, S. J., ABUHALOOB, L., RICHARDS, D., HECTOR, M. P. & FREEMAN, R. 2017. The relationship between childhood body weight and dental caries experience: an umbrella systematic review protocol. *Systematic Reviews,* 6**,** 216.

CASAMASSIMO, P., LEE, J., MARAZITA, M., MILGROM, P., CHI, D. & DIVARIS, K. 2014. Improving Children's Oral Health An Interdisciplinary Research Framework. *Journal of dental research***,** 938-942.

CASTILHO, A. R., MIALHE, F. L., BARBOSA TDE, S. & PUPPIN-RONTANI, R. M. 2013. Influence of family environment on children's oral health: a systematic review. *J Pediatr (Rio J),* 89**,** 116-23.

CASTRO, P. D., KEARNEY, J. & LAYTE, R. 2015. A study of early complementary feeding determinants in the Republic of Ireland based on a cross-sectional analysis of the Growing Up in Ireland infant cohort. *Public Health Nutr,* 18**,** 292-302.

CASTRO, P. D., LAYTE, R. & KEARNEY, J. 2014. Ethnic variation in breastfeeding and complimentary feeding in the Republic of Ireland. *Nutrients,* 6**,** 1832-1849.

CHAFFEE, B. W., FEATHERSTONE, J. D., GANSKY, S. A., CHENG, J. & ZHAN, L. 2016. Caries Risk Assessment Item Importance: Risk Designation and Caries Status in Children under Age 6. *JDR Clinical & Translational Research,* 1**,** 131-142.

CHAFFEE, B. W., FELDENS, C. A., RODRIGUES, P. H. & VITOLO, M. R. 2015. Feeding practices in infancy associated with caries incidence in early childhood. *Community Dent Oral Epidemiol,* 43**,** 338-48.

CHANKANKA, O., LEVY, S. M., MARSHALL, T. A., CAVANAUGH, J. E., WARREN, J. J., BROFFITT, B. & KOLKER, J. L. 2015. The associations between dietary intakes from 36 to 60 months of age and primary dentition non-cavitated caries and cavitated caries. *J Public Health Dent,* 75**,** 265-73.

CHAWLA, N. 2010. *The Data Mining and Knowledge Discovery Handbook*. Maimon O., Rokach L., editors. Berlin: Springer.

CHEN, J. H. & ASCH, S. M. 2017. Machine learning and prediction in medicine-beyond the peak of inflated expectations. *N Engl J Med,* 376**,** 2507-2509.

CHENG, C.-W., MARTIN, G. S., WU, P.-Y. & WANG, M. D. PHARM-Association Rule Mining for Predictive Health. The International Conference on Health Informatics, 2014. Springer, 114-117.

CHI, D. L., LUU, M. & CHU, F. 2017. A scoping review of epidemiologic risk factors for pediatric obesity: Implications for future childhood obesity and dental caries prevention research. *J Public Health Dent,* 77 Suppl 1**,** S8-S31.

CHRISTENSEN, R. & LANGBERG, H. 2012. Statistical principles for prospective study protocols:: design, analysis, and reporting. *International journal of sports physical therapy,* 7**,** 504.

CLEMENS, R. A., JONES, J. M., KERN, M., LEE, S.-Y., MAYHEW, E. J., SLAVIN, J. L. & ZIVANOVIC, S. 2016. Functionality of Sugars in Foods and Health. *Comprehensive Reviews in Food Science and Food Safety,* 15**,** 433-470.

CLEVELAND, W. S. & MCGILL, R. 1985. Graphical perception and graphical methods for analyzing scientific data. *Science,* 229**,** 828-833.

COLE, T. J. 1990. The LMS method for constructing normalized growth standards. *Eur J Clin Nutr,* 44**,** 45-60.

COLE, T. J., BELLIZZI, M. C., FLEGAL, K. M. & DIETZ, W. H. 2000. Establishing a standard definition for child overweight and obesity worldwide: international survey. *BMJ,* 320**,** 1240.

COLE, T. J., FLEGAL, K. M., NICHOLLS, D. & JACKSON, A. A. 2007. Body mass index cut offs to define thinness in children and adolescents: international survey. *BMJ,* 335**,** 194.

COLE, T. J. & LOBSTEIN, T. 2012. Extended international (IOTF) body mass index cut-offs for thinness, overweight and obesity. *Pediatr Obes,* 7**,** 284-294.

COLE, T. J., WRIGHT, C. M., WILLIAMS, A. F. & GROUP, R. G. C. E. 2012. Designing the new UK-WHO growth charts to enhance assessment of growth around birth. *Arch Dis Child Fetal Neonatal Ed,* 97**,** F219-22.

CONDON, J. T. & CORKINDALE, C. J. 1998. The assessment of parent-to-infant attachment: Development of a self-report questionnaire instrument. *Journal of Reproductive and Infant Psychology,* 16**,** 57-76.

CONNOLLY, A., HEARTY, A., NUGENT, A., MCKEVITT, A., BOYLAN, E., FLYNN, A. & GIBNEY, M. J. 2010. Pattern of intake of food additives associated with hyperactivity in Irish children and teenagers. *Food Addit Contam Part A Chem Anal Control Expo Risk Assess,* 27**,** 447-56.

COSTACURTA, M., DIRENZO, L., SICURO, L., GRATTERI, S., DE LORENZO, A. & DOCIMO, R. 2014. Dental caries and childhood obesity: analysis of food intakes, lifestyle. *Eur J Paediatr Dent,* 15**,** 343-8.

CROWE, M., O' SULLIVAN, M., CASSETTI, O. & O' SULLIVAN, A. 2017. Weight Status and Dental Problems in Early Childhood: Classification Tree Analysis of a National Cohort. *Dentistry Journal,* 5**,** 25.

CROWE, M., O'SULLIVAN, A., MCGRATH, C., CASSETTI, O., SWORDS, L. & O'SULLIVAN, M. 2016. Early Childhood Dental Problems Classification Tree Analyses of 2 Waves of an Infant Cohort Study. *JDR Clinical & Translational Research,* 1**,** 275-284.

CURZON, M. & HEFFERREN, J. 2001. Nutrition: modern methods for assessing the cariogenic and erosive potential of foods. *British dental journal,* 191**,** 41-46.

DAHER, A., ABREU, M. H. & COSTA, L. R. 2015. Recognizing preschool children with primary teeth needing dental treatment because of caries-related toothache. *Community Dent Oral Epidemiol,* 43**,** 298-307.

DALGAARD, P. 2008. *Introductory statistics with R,* New York, Springer Science & Business Media.

DARMAWIKARTA, D., CHEN, Y., CARSLEY, S., BIRKEN, C. S., PARKIN, P. C., SCHROTH, R. J. & MAGUIRE, J. L. 2014. Factors Associated With Dental Care Utilization in Early Childhood. *Pediatrics,* 133**,** e1594-e1600.

DE ONIS, M., BLOSSNER, M. & BORGHI, E. 2010. Global prevalence and trends of overweight and obesity among preschool children. *Am J Clin Nutr,* 92**,** 1257-64.

DE ONIS, M. & LOBSTEIN, T. 2010. Defining obesity risk status in the general childhood population: which cut-offs should we use? *Int J Pediatr Obes,* 5**,** 458-60.

DECLERCK, D., LEROY, R., MARTENS, L., LESAFFRE, E., GARCIA-ZATTERA, M.-J., BROUCKE, S. V., DEBYSER, M. & HOPPENBROUWERS, K. 2008. Factors associated with prevalence and severity of caries experience in preschool children. *Community Dentistry and Oral Epidemiology,* 36**,** 168-178.

DEPARTMENT OF CHILDREN AND YOUTH AFFAIRS. 2013. *Right From the Start: Report of The Expert Advisory Group on The Early Years Strategy.* Available: https://www.dcya.gov.ie/documents/policy/RightFromTheStart.pdf[Accessed 22 March 2016.

DEPARTMENT OF CHILDREN AND YOUTH AFFAIRS. 2014. Better Outcomes, Brighter Futures: The National Policy Framework for Children and Young People 2014 - 2020 (Briefing Note). Available:http://lenus.ie/hse/handle/10147/136334/simplesearch?filter_field_0=subject&filter_type_0=equals&filter_value_0=CHILD+HEALTH&sort_by=dateissued&order=DESC [Accessed 3 March 2016].

DIAZ-GARRIDO, N., LOZANO, C. & GIACAMAN, R. A. 2016. Frequency of sucrose exposure on the cariogenicity of a biofilm-caries model. *European journal of dentistry,* 10**,** 345.

DIVARIS, K. 2016. Predicting Dental Caries Outcomes in Children: A "Risky" Concept. *J Dent Res,* 95**,** 248-54.

DIVARIS, K., BHASKAR, V. & MCGRAW, K. A. 2017. Pediatric obesity-related curricular content and training in dental schools and dental hygiene programs: systematic review and recommendations. *J Public Health Dent,* 77 Suppl 1**,** S96-S103.

DO, T., DEVINE, D. & MARSH, P. D. 2013. Oral biofilms: molecular analysis, challenges, and future prospects in dental diagnostics. *Clin Cosmet Investig Dent,* 5**,** 11-9.

DUIJSTER, D., O'MALLEY, L., ELISON, S., VAN LOVEREN, C., MARCENES, W., ADAIR, P. M. & PINE, C. M. 2013a. Family relationships as an explanatory variable in childhood dental caries: a systematic review of measures. *Caries Res,* 47 Suppl 1**,** 22-39.

DUIJSTER, D., VERRIPS, G. H. & VAN LOVEREN, C. 2013b. The role of family functioning in childhood dental caries. *Community Dent Oral Epidemiol*.

DYE, B. A. 2017. The Global Burden of Oral Disease: Research and Public Health Significance. *J Dent Res,* 96**,** 361-363.

DYE, B. A., AREVALO, O. & VARGAS, C. M. 2010. Trends in paediatric dental caries by poverty status in the United States, 1988–1994 and 1999–2004. *International Journal of Paediatric Dentistry,* 20**,** 132-143.

DYE, B. A., HSU, K. L. & AFFUL, J. 2015. Prevalence and Measurement of Dental Caries in Young Children. *Pediatr Dent,* 37**,** 200-16.

DYE, B. A., SHENKIN, J. D., OGDEN, C. L., MARSHALL, T. A., LEVY, S. M. & KANELLIS, M. J. 2004. The relationship between healthful eating practices and dental caries in children aged 2–5 years in the United States, 1988–1994. *The Journal of the American Dental Association,* 135**,** 55-66.

EDWARDS, C. H., ROSSI, M., CORPE, C. P., BUTTERWORTH, P. J. & ELLIS, P. R. 2016. The role of sugars and sweeteners in food, diet and health: Alternatives for the future. *Trends in Food Science & Technology,* 56**,** 158-166.

ERICKSON, J., SADEGHIRAD, B., LYTVYN, L., SLAVIN, J. & JOHNSTON, B. C. 2017. The Scientific Basis of Guideline Recommendations on Sugar Intake: A Systematic Review. *Ann Intern Med,* 166**,** 257-267.

ERICKSON, J. & SLAVIN, J. 2015a. Are restrictive guidelines for added sugars science based? *Nutr J,* 14**,** 124.

ERICKSON, J. & SLAVIN, J. 2015b. Total, added, and free sugars: are restrictive guidelines science-based or achievable? *Nutrients,* 7**,** 2866-78.

FABER, M., WENHOLD, F. A., MACINTYRE, U. E., WENTZEL-VILJOEN, E., STEYN, N. P. & OLDEWAGE-THERON, W. H. 2013. Presentation and interpretation of food intake data: factors affecting comparability across studies. *Nutrition,* 29**,** 1286-92.

FARAJIAN, P., RISVAS, G., PANAGIOTAKOS, D. B. & ZAMPELAS, A. 2016. Food sources of free sugars in children's diet and identification of lifestyle patterns associated with free sugars intake: the GRECO (Greek Childhood Obesity) study. *Public Health Nutr,* 19**,** 2326-2335.

FARPOUR-LAMBERT, N. J., BAKER, J. L., HASSAPIDOU, M., HOLM, J. C., NOWICKA, P., O'MALLEY, G. & WEISS, R. 2015. Childhood Obesity Is a Chronic Disease Demanding Specific Health Care--a Position Statement from the Childhood Obesity Task Force (COTF) of the European Association for the Study of Obesity (EASO). *Obes Facts,* 8**,** 342-9.

FINLAYSON, T. L., SIEFERT, K., ISMAIL, A. I. & SOHN, W. 2007. Psychosocial factors and early childhood caries among low-income African–American children in Detroit. *Community Dentistry and Oral Epidemiology,* 35**,** 439-448.

FISHER-OWENS, S. A., GANSKY, S. A., PLATT, L. J., WEINTRAUB, J. A., SOOBADER, M. J., BRAMLETT, M. D. & NEWACHECK, P. W. 2007.

Influences on children's oral health: a conceptual model. *Pediatrics,* 120**,** e510-20.

FISHER-OWENS, S. A., ISONG, I. A., SOOBADER, M. J., GANSKY, S. A., WEINTRAUB, J. A., PLATT, L. J. & NEWACHECK, P. W. 2013. An examination of racial/ethnic disparities in children's oral health in the United States. *J Public Health Dent,* 73**,** 166-174.

FISHER, J. O., WRIGHT, G., HERMAN, A. N., MALHOTRA, K., SERRANO, E. L., FOSTER, G. D. & WHITAKER, R. C. 2015. "Snacks are not food". Low-income, urban mothers' perceptions of feeding snacks to their preschool-aged children. *Appetite,* 84**,** 61-7.

FLEGAL, K. M. & COLE, T. J. 2013. *Construction of LMS parameters for the Centers for Disease Control and Prevention 2000 growth charts*, US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

FLEGAL, K. M. & OGDEN, C. L. 2011. Childhood obesity: are we all speaking the same language? *Adv Nutr,* 2**,** 159S-66S.

FOOD SAFETY AUTHORITY OF IRELAND. 2011a. *Scientific Recommendations for a National Infant Feeding Policy, 2nd edn* [Online]. Dublin, Ireland: Food Safety Authority of Ireland. Available: https://www.fsai.ie/science_and_health/healthy_eating.html [Accessed 9 February 2017].

FOOD SAFETY AUTHORITY OF IRELAND. 2011b. *Scientific Recommendations for Healthy Eating Guidelines in Ireland* [Online]. Dublin, Ireland: Food Safety Authority of Ireland. Available:https://www.fsai.ie/science_and_health/healthy_eating.html[Accessed 4 February 2017].

FOOD STANDARDS AUSTRALIA NEW ZEALAND. 2014. *AUSNUT 2011-2013- food composition database* Available:http://www.foodstandards.gov.au/science/monitoringnutrients/ausnut/foodnutrient/Pages/default.aspx. [Accessed 11 May 2017].

FORSTER, H., FALLAIZE, R., GALLAGHER, C., O'DONOVAN, C. B., WOOLHEAD, C., WALSH, M. C., MACREADY, A. L., LOVEGROVE, J. A., MATHERS, J. C. & GIBNEY, M. J. 2014. Online dietary intake estimation: the Food4Me food frequency questionnaire. *Journal of medical Internet research,* 16.

FOSTER, E. & ADAMSON, A. 2014. Challenges involved in measuring intake in early life: focus on methods. *Proc Nutr Soc,* 73**,** 201-9.

FRIEL, S., NOLAN, G. & KELLEHER, C. 1996. *Changes in the food chain since the time of the Great Irish Famine.* National Nutrition Surveillance Centre. Available: http://www.ucd.ie/t4cms/nnsc_MHeinen_position%20paper%20series 3.pdf [Accessed 20 April 2015].

GARBARINO, J. 1982. *Children and families in the social environment,* New York, Transaction Publishers.

GIBNEY, M. J. & WOLEVER, T. M. S. 1997. Periodicity of eating and human health: present perspective and future directions. *British Journal of Nutrition,* 77**,** S3.

GIBNEY, M. J. & WOLMARANS, P. 2004. Dietary Guidelines. *In:* GIBNEY, M. J., MARGETTS, B. M., KEARNEY, J. M. & AND ARAB, L. (eds.) *Public Health Nutr.* Oxford, UK: Blackwell Publishing, 133-143.

GIBSON, E. L., KREICHAUF, S., WILDGRUBER, A., VÖGELE, C., SUMMERBELL, C., NIXON, C., MOORE, H., DOUTHWAITE, W. & MANIOS, Y. 2012. A narrative review of psychological and educational strategies applied to young children's eating behaviours aimed at reducing obesity risk. *Obesity reviews,* 13**,** 85-95.

GIBSON, S., FRANCIS, L., NEWENS, K. & LIVINGSTONE, B. 2016. Associations between free sugars and nutrient intakes among children and adolescents in the UK. *Br J Nutr,* 116**,** 1265-1274.

GIBSON, S. & WILLIAMS, S. 1999. Dental caries in pre-school children: Associations with social class, toothbrushing habit and consumption of sugars and sugar-containing foods: Further analysis of data. *Caries Res,* 33**,** 101-13.

GLICK, M., WILLIAMS, D. M., KLEINMAN, D. V., VUJICIC, M., WATT, R. G. & WEYANT, R. J. 2017. A new definition for oral health developed by the FDI World Dental Federation opens the door to a universal definition of oral health. *J Am Dent Assoc,* 147**,** 915-917.

GOLLEY, R. K., BELL, L. K., HENDRIE, G. A., RANGAN, A. M., SPENCE, A., MCNAUGHTON, S. A., CARPENTER, L., ALLMAN-FARINELLI, M., DE SILVA, A., GILL, T., COLLINS, C. E., TRUBY, H., FLOOD, V. M. & BURROWS, T. 2017. Validity of short food questionnaire items to measure intake in children and adolescents: a systematic review. *J Hum Nutr Diet,* 30**,** 36-50.

GOMEZ, A., ESPINOZA, J. L., HARKINS, D. M., LEONG, P., SAFFERY, R., BOCKMANN, M., TORRALBA, M., KUELBS, C., KODUKULA, R. & INMAN, J. 2017. Host Genetic Control of the Oral Microbiome in Health and Disease. *Cell Host & Microbe,* 22**,** 269-278. e3.

GOODMAN, R. 1997. The Strengths and Difficulties Questionnaire: a research note. *Journal of child psychology and psychiatry,* 38**,** 581-586.

GOODSON, J. M., TAVARES, M., WANG, X., NIEDERMAN, R., CUGINI, M., HASTURK, H., BARAKE, R., ALSMADI, O., AL-MUTAWA, S. & ARIGA, J. 2013. Obesity and Dental Decay: Inference on the Role of Dietary Sugar. *PLoS One,* 8**,** e74461.

GOODWIN, M., PATEL, D., VYAS, A., KHAN, A., MCGRADY, M., BOOTHMAN, N. & PRETTY, I. 2017. Sugar before bed: a simple dietary risk factor for caries experience. *Community Dent Health,* 34**,** 8-13.

GUSSY, M., ASHBOLT, R., CARPENTER, L., VIRGO-MILTON, M., CALACHE, H., DASHPER, S., LEONG, P., DE SILVA, A., DE LIVERA, A., SIMPSON, J. & WATERS, E. 2016. Natural history of dental caries in very young Australian children. *Int J Paediatr Dent,* 26**,** 173-83.

GUSSY, M. G., WATERS, E. G., WALSH, O. & KILPATRICK, N. M. 2006. Early childhood caries: Current evidence for aetiology and prevention. *J Paediatr Child Health,* 42**,** 37-43.

GUSTAFSSON, B., QUENSEL, C. & SWENANDER, L. 1954. L., Lundqvist, C., Grahne, H., Bonow, BE and Krasse, B.: The Vipeholm dental caries study The effect of different levels of carbohydrate intake on caries activity in 436 individuals observed for five years. *Acta odontol. scand,* 11**,** 232-364.

HALEVY, A., NORVIG, P. & PEREIRA, F. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems,* 24**,** 8-12.

HALL, K. D., BEMIS, T., BRYCHTA, R., CHEN, K. Y., COURVILLE, A., CRAYNER, E. J., GOODWIN, S., GUO, J., HOWARD, L., KNUTH, N. D., MILLER, B. V., 3RD, PRADO, C. M., SIERVO, M., SKARULIS, M. C., WALTER, M., WALTER, P. J. & YANNAI, L. 2015. Calorie for Calorie, Dietary Fat Restriction Results in More Body Fat Loss than Carbohydrate Restriction in People with Obesity. *Cell Metab,* 22**,** 427-36.

HARRIS, R., NICOLL, A. D., ADAIR, P. M. & PINE, C. M. 2004. Risk factors for dental caries in young children: a systematic review of the literature. *Community Dent Health,* 21**,** 71-85.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Biometrics*.

HAUSEN, H. 1997. Caries prediction–state of the art. *Community Dentistry and Oral Epidemiology,* 25**,** 87-96.

HAYDEN, C., BOWLER, J. O., CHAMBERS, S., FREEMAN, R., HUMPHRIS, G., RICHARDS, D. & CECIL, J. E. 2012. Obesity and dental caries in children: a systematic review and meta-analysis. *Community Dent Oral Epidemiol*.

HEYMAN, M. B. & ABRAMS, S. A. 2017. Fruit Juice in Infants, Children, and Adolescents: Current Recommendations. *Pediatrics***,** e20170967.

HO YU, C. 2010. Exploratory data analysis in the context of data mining and resampling. *International Journal of Psychological Research,* 3.

HONG, L., AHMED, A., MCCUNNIFF, M., OVERMAN, P. & MATHEW, M. 2008. Obesity and Dental Caries in Children Aged 2-6 Years in the

United States: National Health and Nutrition Examination Survey 1999-2002. *J Public Health Dent,* 68**,** 227-233.

HOOLEY, M., SKOUTERIS, H., BOGANIN, C., SATUR, J. & KILPATRICK, N. 2012a. Body mass index and dental caries in children and adolescents: a systematic review of literature published 2004 to 2011. *Syst Rev,* 1**,** 57.

HOOLEY, M., SKOUTERIS, H., BOGANIN, C., SATUR, J. & KILPATRICK, N. 2012b. Parental influence and the development of dental caries in children aged 0–6 years: A systematic review of the literature. *J Dent,* 40**,** 873-885.

HOOLEY, M., SKOUTERIS, H. & MILLAR, L. 2012c. The relationship between childhood weight, dental caries and eating practices in children aged 4-8 years in Australia, 2004-2008. *Pediatr Obes,* 7**,** 461-70.

HOPCRAFT, M. S. & BEAUMONT, S. 2016. The growing problems of dental caries and obesity: an Australian perspective. *Br Dent J,* 221**,** 379-381.

HÖPPNER, F. 2005. Association rules. *Data Mining and Knowledge Discovery Handbook.* Springer, 353-376.

HU, F. B. 2013. Resolved: there is sufficient scientific evidence that decreasing sugar-sweetened beverage consumption will reduce the prevalence of obesity and obesity-related diseases. *Obes Rev,* 14**,** 606-19.

HUJOEL, P. P. & LINGSTROM, P. 2017. Nutrition, dental caries and periodontal disease: a narrative review. *J Clin Periodontol,* 44 Suppl 18**,** S79-S84.

INDURKHYA, N. & DAMERAU, F. J. 2010. *Handbook of natural language processing,* Florida, CRC Press.

INSELBERG, A. 1985. The plane with parallel coordinates. *The visual computer,* 1**,** 69-91.

IOANNIDIS, J. P. 2013. Implausible results in human nutrition research. *BMJ : British Medical Journal,* 347.

IOANNIDIS, J. P. & TREPANOWSKI, J. F. 2018. Disclosures in Nutrition Research: Why It Is Different. *JAMA,* 319**,** 547-548.

IRISH UNIVERSITIES NUTRITION ALLIANCE. 2012. *National Preschool Nutrition Survey 2010-11* [Online]. Available: http://www.iuna.net/?p=169 [Accessed November 2016].

ISMAIL, A., SOHN, W., LIM, S. & WILLEM, J. 2009. Predictors of dental caries progression in primary teeth. *Journal of dental research,* 88**,** 270-275.

ISMAIL, A. I. 2003. Determinants of health in children and the problem of early childhood caries. *Pediatr Dent,* 25**,** 328-333.

JACQUIER, E. F., GATRELL, A. & BINGLEY, A. 2017. "We don't snack": Attitudes and perceptions about eating in-between meals amongst caregivers of young children. *Appetite,* 108**,** 483-490.

JAMES, G., WITTEN, D., HASTIE, T. & TIBSHIRANI, R. 2013. *An introduction to statistical learning,* New York, Springer.

JENAB, M., SLIMANI, N., BICTASH, M., FERRARI, P. & BINGHAM, S. A. 2009. Biomarkers in nutritional epidemiology: applications, needs and new horizons. *Human genetics,* 125**,** 507-525.

JIN, L. J., LAMSTER, I. B., GREENSPAN, J. S., PITTS, N. B., SCULLY, C. & WARNAKULASURIYA, S. 2016. Global burden of oral diseases: emerging concepts, management and interplay with systemic health. *Oral Dis,* 22**,** 609-19.

JOHANSSON, I., HOLGERSON, P. L., KRESSIN, N. R., NUNN, M. E. & TANNER, A. C. 2010. Snacking habits and caries in young children. *Caries Res,* 44**,** 421-30.

JOHNSON, S., CARPENTER, L., AMEZDROZ, E., DASHPER, S., GUSSY, M., CALACHE, H., DE SILVA, A. M. & WATERS, E. 2016. Cohort Profile: The VicGeneration (VicGen) study: An Australian oral health birth cohort. *Int J Epidemiol,* 46**,** 29-30g.

JOYCE, T., MCCARTHY, S. N. & GIBNEY, M. J. 2008. Relationship between energy from added sugars and frequency of added sugars intake in Irish children, teenagers and adults. *Br J Nutr,* 99**,** 1117-26.

KAHN, R. & SIEVENPIPER, J. L. 2014. Dietary sugar and body weight: Have we reached a crisis in the epidemic of obesity and diabetes? *Diabetes Care,* 37**,** 957-962.

KAMPSTRA, P. 2008. Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of Statistical Software,* 28**,** CS1.

KANTOVITZ, K. R., PASCON, F. M., RONTANI, R. M. P. & GAVIAO, M. B. D. 2006. Obesity and dental caries-A systematic review. *Oral Health and Preventive Dentistry,* 4**,** 137.

KARLSEN, S., MORRIS, S., KINRA, S., VALLEJO-TORRES, L. & VINER, R. M. 2014. Ethnic variations in overweight and obesity among children over time: findings from analyses of the Health Surveys for England 1998-2009. *Pediatr Obes,* 9**,** 186-96.

KASS, G. V. 1980. An exploratory technique for investigating large quantities of categorical data. *Applied statistics***,** 119-127.

KASSEBAUM, N. J., BERNABE, E., DAHIYA, M., BHANDARI, B., MURRAY, C. J. & MARCENES, W. 2015. Global burden of untreated caries: a systematic review and metaregression. *J Dent Res,* 94**,** 650-8.

KEANE, E., KEARNEY, P. M., PERRY, I. J., KELLEHER, C. C. & HARRINGTON, J. M. 2014. Trends and prevalence of overweight and

obesity in primary school aged children in the Republic of Ireland from 2002-2012: a systematic review. *BMC Public Health,* 14**,** 974.

KETTLER, S., KENNEDY, M., MCNAMARA, C., OBERDORFER, R., O'MAHONY, C., SCHNABEL, J., SMITH, B., SPRONG, C., FALUDI, R. & TENNANT, D. 2015. Assessing and reporting uncertainties in dietary exposure analysis: Mapping of uncertainties in a tiered approach. *Food Chem Toxicol,* 82**,** 79-95.

KHOURY, M. J., LAM, T. K., IOANNIDIS, J. P., HARTGE, P., SPITZ, M. R., BURING, J. E., CHANOCK, S. J., CROYLE, R. T., GODDARD, K. A., GINSBURG, G. S., HERCEG, Z., HIATT, R. A., HOOVER, R. N., HUNTER, D. J., KRAMER, B. S., LAUER, M. S., MEYERHARDT, J. A., OLOPADE, O. I., PALMER, J. R., SELLERS, T. A., SEMINARA, D., RANSOHOFF, D. F., REBBECK, T. R., TOURASSI, G., WINN, D. M., ZAUBER, A. & SCHULLY, S. D. 2013. Transforming epidemiology for 21st century medicine and public health. *Cancer Epidemiol Biomarkers Prev,* 22**,** 508-16.

KIM, J. K. & RAO, J. N. K. 2011. Combining data from two independent surveys: a model-assisted approach. *Biometrika,* 99**,** 85-100.

KIM SEOW, W. 2012. Environmental, maternal, and child factors which contribute to early childhood caries: a unifying conceptual model. *International Journal of Paediatric Dentistry,* 22**,** 157-168.

KINGSFORD, C. & SALZBERG, S. L. 2008. What are decision trees? *Nat Biotechnol,* 26**,** 1011-3.

KIRKPATRICK, S. I., REEDY, J., BUTLER, E. N., DODD, K. W., SUBAR, A. F., THOMPSON, F. E. & MCKINNON, R. A. 2014. Dietary assessment in food environment research: a systematic review. *Am J Prev Med,* 46**,** 94-102.

KREBS-SMITH, S. M., SUBAR, A. F. & REEDY, J. 2015. Examining Dietary Patterns in Relation to Chronic Disease: Matching Measures and Methods to Questions of Interest. *Circulation,* 132**,** 790-3.

KRIEGER, N. 2008. Proximal, distal, and the politics of causation: what's level got to do with it? *Am J Public Health,* 98**,** 221-30.

KRIEGER, N. 2012. Who and What Is a 'Population'? Historical Debates, Current Controversies, and Implications for Understanding 'Population Health' and Rectifying Health Inequities. *Milbank Quarterly,* 90**,** 634-681.

KUH, D., BEN-SHLOMO, Y., LYNCH, J., HALLQVIST, J. & POWER, C. 2003. Life course epidemiology. *J Epidemiol Community Health,* 57**,** 778.

KUHN, L., PAGE, K., WARD, J. & WORRALL-CARTER, L. 2014. The process and utility of classification and regression tree methodology in nursing research. *J Adv Nurs,* 70**,** 1276-86.

LAKENS, D. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science,* 8**,** 355-362.

LANDER, J. P. 2014. *R for everyone: Advanced analytics and graphics,* New Jersey, Pearson Education.

LAVALLE, P. S., GLAROS, A., BOHATY, B. & MCCUNNIFF, M. 2000. The effect of parental stress on the oral health of children. *Journal of Clinical Psychology in Medical Settings,* 7**,** 197-201.

LAYTE, R., BENNETT, A., MCCRORY, C. & KEARNEY, J. 2014a. Social class variation in the predictors of rapid growth in infancy and obesity at age 3 years. *International Journal of Obesity,* 38**,** 82.

LAYTE, R., BENNETT, A., MCCRORY, C. & KEARNEY, J. 2014b. Social class variation in the predictors of rapid growth in infancy and obesity at age 3 years. *International Journal of Obesity,* 38**,** 82-90.

LEE, J. K., WILLIAMS, P. D. & CHEON, S. 2008. Data mining in genomics. *Clinics in laboratory medicine,* 28**,** 145-166.

LEE, J. Y. & DIVARIS, K. 2014. The ethical imperative of addressing oral health disparities: a unifying framework. *J Dent Res,* 93**,** 224-30.

LEE, J. Y., WATT, R. G., WILLIAMS, D. M. & GIANNOBILE, W. V. 2016. A New Definition for Oral Health: Implications for Clinical Practice, Policy, and Research. *J Dent Res,* 96**,** 125-127.

LEE, P. H., MCGRATH, C. P., KONG, A. Y. & LAM, T. H. 2013. Self-report poor oral health and chronic diseases: the Hong Kong FAMILY project. *Community Dent Oral Epidemiol,* 41**,** 451-458.

LEECH, R. M., WORSLEY, A., TIMPERIO, A. & MCNAUGHTON, S. A. 2015. Understanding meal patterns: definitions, methodology and impact on nutrient intake and diet quality. *Nutrition research reviews,* 28**,** 1-21.

LEEK, J. T. & PENG, R. D. 2015a. Statistics: P values are just the tip of the iceberg. *Nature,* 520**,** 612.

LEEK, J. T. & PENG, R. D. 2015b. What is the question? *Science,* 347**,** 1314-1315.

LEI, L., RANGAN, A., FLOOD, V. M. & LOUIE, J. C. 2016. Dietary intake and food sources of added sugar in the Australian population. *Br J Nutr,* 115**,** 868-77.

LEMON, S. C., ROY, J., CLARK, M. A., FRIEDMANN, P. D. & RAKOWSKI, W. 2003. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of behavioral medicine,* 26**,** 172-181.

LEONG, P. M., GUSSY, M. G., BARROW, S. Y. L., SILVA-SANIGORSKI, A. & WATERS, E. 2013. A systematic review of risk factors during first year

of life for early childhood caries. *International Journal of Paediatric Dentistry,* 23**,** 235-250.

LEROY, R., BOGAERTS, K., HOPPENBROUWERS, K., MARTENS, L. C. & DECLERCK, D. 2013. Dental attendance in preschool children - a prospective study. *Int J Paediatr Dent,* 23**,** 84-93.

LESKOVEC, J., RAJARAMAN, A. & ULLMAN, J. D. 2014. *Mining of massive datasets*, Cambridge university press.

LIANG, J. J., ZHANG, Z. Q., CHEN, Y. J., MAI, J. C., MA, J., YANG, W. H. & JING, J. 2016. Dental caries is negatively correlated with body mass index among 7-9 years old children in Guangzhou, China. *BMC Public Health,* 16**,** 638.

LINGSTRÖM, P. 2009. Impact of food sugars and polysaccharides on dental caries. *Food Constituents and Oral Health: Current Status and Future Prospects***,** 163.

LINGSTROM, P., VAN HOUTE, J. & KASHKET, S. 2000. Food Starches and Dental Caries. *Critical Reviews in Oral Biology & Medicine,* 11**,** 366-380.

LISSNER, L. 2006. Measuring food intake in studies of obesity. *Public Health Nutr,* 5**,** 889-892.

LIVINGSTONE, M. B. & BLACK, A. E. 2003. Markers of the validity of reported energy intake. *J Nutr,* 133 Suppl 3**,** 895s-920s.

LLENA, C. & FORNER, L. 2008. Dietary habits in a child population in relation to caries experience. *Caries Res,* 42**,** 387-93.

LO, J. C., MARING, B., CHANDRA, M., DANIELS, S. R., SINAIKO, A., DALEY, M. F., SHERWOOD, N. E., KHARBANDA, E. O., PARKER, E. D., ADAMS, K. F., PRINEAS, R. J., MAGID, D. J., O'CONNOR, P. J. & GREENSPAN, L. C. 2014. Prevalence of obesity and extreme obesity in children aged 3-5 years. *Pediatr Obes,* 9**,** 167-75.

LOBSTEIN, T. & DAVIES, S. 2009. Defining and labelling 'healthy' and 'unhealthy' food. *Public Health Nutr,* 12**,** 331-40.

LOBSTEIN, T., JACKSON-LEACH, R., MOODIE, M. L., HALL, K. D., GORTMAKER, S. L., SWINBURN, B. A., JAMES, W. P. T., WANG, Y. & MCPHERSON, K. 2015. Child and adolescent obesity: part of a bigger picture. *The Lancet,* 385**,** 2510-2520.

LOH, W.-Y. 2014. Fifty Years of Classification and Regression Trees. *International Statistical Review,* 82**,** 329-348.

LOH, W.-Y. & SHIH, Y.-S. 1997. Split selection methods for classification trees. *Statistica sinica***,** 815-840.

LOUIE, J. C., MOSHTAGHIAN, H., BOYLAN, S., FLOOD, V. M., RANGAN, A. M., BARCLAY, A. W., BRAND-MILLER, J. C. & GILL, T. P. 2015. A

systematic methodology to estimate added sugar content of foods. *Eur J Clin Nutr,* 69**,** 154-61.

LUKE, D. A. & STAMATAKIS, K. A. 2012. Systems science methods in public health: dynamics, networks, and agents. *Annu Rev Public Health,* 33**,** 357-76.

LUSTIG, R. H., SCHMIDT, L. A. & BRINDIS, C. D. 2012. Public health: The toxic truth about sugar. *Nature,* 482**,** 27-29.

LUZZI, L., CHRISOPOULOS, S. & BRENNAN, D. S. 2013. Decline in usually visiting the dentist for a problem in Australia, 1994 to 2010: an age–period–cohort analysis. *Community Dent Oral Epidemiol,* 42**,** 349-357.

MACKERRAS, D. & MARGETTS, B. M. 2005. Nutritional Epidemiology. *Handbook of Epidemiology.* 999-1042.

MACLEOD, M. 2018. What makes interdisciplinarity difficult? Some consequences of domain specificity in interdisciplinary practice. *Synthese,* 195**,** 697-720.

MAGAREY, A., WATSON, J., GOLLEY, R. K., BURROWS, T., SUTHERLAND, R., MCNAUGHTON, S. A., DENNEY-WILSON, E., CAMPBELL, K. & COLLINS, C. 2011. Assessing dietary intake in children and adolescents: Considerations and recommendations for obesity research. *Int J Pediatr Obes,* 6**,** 2-11.

MAIMON, O. & ROKACH, L. 2005. Classification Trees. *In:* MAIMON, O. & ROKACH, L. (eds.) *Data mining and knowledge discovery handbook.* New York: Springer, 149-174.

MAIMON, O. & ROKACH, L. 2009. Introduction to knowledge discovery and data mining. *Data Mining and Knowledge Discovery Handbook.* New York: Springer, 1-15.

MAINDONALD, J. & BRAUN, W. J. 2010. *Data Analysis and Graphics Using R: An Example-Based Approach*, Cambridge University Press.

MÄNTYMAA, M., PUURA, K., LUOMA, I., SALMELIN, R. K. & TAMMINEN, T. 2006. Mother's early perception of her infant's difficult temperament, parenting stress and early mother-infant interaction. *Nordic journal of psychiatry,* 60**,** 379-386.

MARSH, P. D. 2006. Dental plaque as a biofilm and a microbial community - implications for health and disease. *BMC Oral Health,* 6 Suppl 1**,** S14.

MARSHALL, T. 2013. Preventing dental caries associated with sugar-sweetened beverages. *Journal of the American Dental Association,* 144**,** 1148-1152.

MARSHALL, T. A. 2015. Nomenclature, characteristics, and dietary intakes of sugars. *The Journal of the American Dental Association,* 146**,** 61-64.

MARSHALL, T. A., BROFFITT, B., EICHENBERGER-GILMORE, J., WARREN, J. J., CUNNINGHAM, M. A. & LEVY, S. M. 2005. The roles of meal, snack, and daily total food and beverage exposures on caries experience in young children. *J Public Health Dent,* 65**,** 166-173.

MARSHALL, T. A., EICHENBERGER-GILMORE, J. M., LARSON, M. A., WARREN, J. J. & LEVY, S. M. 2007a. Comparison of the intakes of sugars by young children with and without dental caries experience. *The Journal of the American Dental Association,* 138**,** 39-46.

MARSHALL, T. A., EICHENBERGER-GILMORE, J. M., BROFFITT, B. A., WARREN, J. J. & LEVY, S. M. 2007b. Dental caries and childhood obesity: roles of diet and socioeconomic status. *Community Dentistry and Oral Epidemiology,* 35**,** 449-458.

MARSHALL, T. A., LEVY, S. M., BROFFITT, B., WARREN, J. J., EICHENBERGER-GILMORE, J. M., BURNS, T. L. & STUMBO, P. J. 2003. Dental caries and beverage consumption in young children. *Pediatrics,* 112**,** e184-e191.

MCCANCE, R. 2002. McCance and Widdowson's The Composition of Foods. compiled by the Food Standards Agency and Institute of Food Research. *Royal Society of Chemistry, Cambridge***,** 537.

MENNELLA, J. A. 2014. Ontogeny of taste preferences: basic biology and implications for health. *Am J Clin Nutr,* 99**,** 704S-11S.

MENON, I., NAGARAJAPPA, R., RAMESH, G. & TAK, M. 2013. Parental stress as a predictor of early childhood caries among preschool children in India. *International Journal of Paediatric Dentistry,* 23**,** 160-165.

MEURMAN, P. K. & PIENIHÄKKINEN, K. 2011. Factors Associated with Caries Increment: A Longitudinal Study from 18 Months to 5 Years of Age. *Caries Res,* 44**,** 519-24.

MEYER, B. D. & LEE, J. Y. 2015. The Confluence of Sugar, Dental Caries, and Health Policy. *J Dent Res,* 94**,** 1338-40.

MILLER, B., FRIDLINE, M., LIU, P. Y. & MARINO, D. 2014. Use of CHAID decision trees to formulate pathways for the early detection of metabolic syndrome in young adults. *Comput Math Methods Med,* 2014**,** 242717.

MINTZ, S. W. 1985. *Sweetness and power: the place of sugar in modern history,* New York, Penguin.

MIS, N. F., BRAEGGER, C., BRONSKY, J., CAMPOY, C., DOMELLÖF, M., EMBLETON, N. D., HOJSAK, I., HULST, J., INDRIO, F. & LAPILLONNE, A. 2017. Sugar in infants, children and adolescents: a position paper of the European Society for Paediatric Gastroenterology, Hepatology and Nutrition Committee on Nutrition. *Journal of pediatric gastroenterology and nutrition,* 65**,** 681-696.

MOIMAZ, S. A., FADEL, C. B., LOLLI, L. F., GARBIN, C. A., GARBIN, A. J. & SALIBA, N. A. 2014. Social aspects of dental caries in the context of mother-child pairs. *J Appl Oral Sci,* 22**,** 73-8.

MOYNIHAN, P. 2016. Sugars and Dental Caries: Evidence for Setting a Recommended Threshold for Intake. *Advances in Nutrition: An International Review Journal,* 7**,** 149-156.

MOYNIHAN, P., MAKINO, Y., PETERSEN, P. E. & OGAWA, H. 2018. Implications of WHO Guideline on Sugars for dental health professionals. *Community Dent Oral Epidemiol,* 46**,** 1-7.

MOYNIHAN, P. & PETERSEN, P. E. 2004. Diet, nutrition and the prevention of dental diseases. *Public Health Nutr,* 7.

MOYNIHAN, P., THOMASON, M., WALLS, A., GRAY-DONALD, K., MORAIS, J. A., GHANEM, H., WOLLIN, S., ELLIS, J., STEELE, J., LUND, J. & FEINE, J. 2009. Researching the impact of oral health on diet and nutritional status: methodological issues. *J Dent,* 37**,** 237-49.

MOYNIHAN, P. J. 2002. Dietary advice in dental practice. *British dental journal,* 193**,** 563-568.

MOYNIHAN, P. J. & KELLY, S. A. M. 2014. Effect on Caries of Restricting Sugars Intake: Systematic Review to Inform WHO Guidelines. *Journal of dental research,* 93**,** 8-18.

MURRAY, A., QUAIL, A., MCCRORY, C. & WILLIAMS, J. 2013. *A Summary Guide to Wave 2 of the Infant Cohort (at 3 years) of Growing Up in Ireland*. Dublin, Ireland: The Economic and Social Research Institute.

NAVIA, J. M. 1994. Carbohydrates and dental health. *Am J Clin Nutr,* 59**,** 719S-727S.

NEWENS, K. J. & WALTON, J. 2016. A review of sugar consumption from nationally representative dietary surveys across the world. *J Hum Nutr Diet,* 29**,** 225-40.

NEWTON, J. T. & BOWER, E. J. 2005. The social determinants of oral health: new approaches to conceptualizing and researching complex causal networks. *Community Dentistry and Oral Epidemiology,* 33**,** 25-34.

NICOLAU, B., MARCENES, W., BARTLEY, M. & SHEIHAM, A. 2003. A life course approach to assessing causes of dental caries experience: the relationship between biological, behavioural, socio-economic and psychological conditions and caries in adolescents. *Caries Res,* 37**,** 319-26.

NICOLAU, B., THOMSON, W. M., STEELE, J. G. & ALLISON, P. J. 2007. Life-course epidemiology: concepts and theoretical models and its relevance to chronic oral conditions. *Community Dent Oral Epidemiol,* 35**,** 241-249.

O'MULLANE, D., JAMES, P., WHELTON, H. & PARNELL, C. 2012. Methodological issues in oral health research: intervention studies. *Community Dent Oral Epidemiol,* 40 Suppl 1**,** 15-20.

O'SULLIVAN, A., GIBNEY, M. J. & BRENNAN, L. 2011. Dietary intake patterns are reflected in metabolomic profiles: potential role in dietary assessment studies. *Am J Clin Nutr,* 93**,** 314-21.

O'CONNELL, A. C. & HARDING, M. 2017. The Republic of Ireland *In:* FOLAYAN, M. O. (ed.) *A Compendium on Oral Health of Children around the World: Early Childhood Caries.* New York: Nova Science Publishers Inc.

ONIS, M. 2006. WHO Child Growth Standards based on length/height, weight and age. *Acta paediatrica,* 95**,** 76-85.

PATEL, R. 2012. *The state of oral health in Europe*. Report commissioned by the platform for better oral health in Europe, 2012.

PENG, R. 2012. *Exploratory data analysis with R*. Available: https://leanpub.com/exdata [Accessed 14 February 2015].

PENG, R. D. & MATSUI, E. 2015. *The Art of Data Science*. *A Guide for Anyone Who Works with Data. Skybrude Consulting* [Online]. Available: http://www.lulu.com/ie/en/shop/roger-peng-and-elizabeth-matsui/the-art-of-data-science/paperback/product-22743811.html[Accessed 30 October 2015].

PEREIRA, M. 2006. The possible role of sugar-sweetened beverages in obesity etiology: a review of the evidence. *International Journal of Obesity,* 30**,** S28-S36.

PERES, M. A., SHEIHAM, A., LIU, P., DEMARCO, F. F., SILVA, A. E., ASSUNCAO, M. C., MENEZES, A. M., BARROS, F. C. & PERES, K. G. 2016. Sugar Consumption and Changes in Dental Caries from Childhood to Adolescence. *J Dent Res,* 95**,** 388-94.

PETERSEN, P. E. & KWAN, S. 2011. Equity, social determinants and public health programmes--the case of oral health. *Community Dent Oral Epidemiol,* 39**,** 481-7.

PETERSON 1963. The Role of Fermentable Carbohydrates in the Production of Dental Caries. *J Public Health Dent,* 23.

PINE, C. 2013. Caring for children's developing mouths. Foreword. *Int Dent J,* 63 Suppl 2**,** 1-2.

PRIEBE, M. G. & MCMONAGLE, J. R. 2016. Effects of Ready-to-Eat-Cereals on Key Nutritional and Health Outcomes: A Systematic Review. *PLoS One,* 11**,** e0164931.

PRIOR, M. R., SANSON, A., SMART, D. & OBERKLAID, F. 2000. *Pathways from infancy to adolescence: Australian Temperament Project 1983-2000*, Australian Institute of Family Studies Melbourne.

PUBLIC HEALTH ENGLAND. 2013a. Dental Epidemiology Programme for England: *Oral Health survey of three-year-old children 2013. A report on the prevalence and severity of dental decay*. Available:http://www.nwph.net/dentalhealth/survey-results5.aspx?id=1[Accessed November 20th, 2013].

PUBLIC HEALTH ENGLAND. 2013b. National Dental Epidemiology Programme for England: *Oral Health survey of five-year-old children 2012. A report on the prevalence and severity of dental decay*. Available:http://www.nwph.net/dentalhealth/survey-results5.aspx?id=1[Accessed 1 November 2013].

PUBLIC HEALTH ENGLAND. 2015a. *The relationship between dental caries and obesity in children:an evidence summary*. Available:https://www.gov.uk/government/publications/dental-caries-and-obesity-their-relationship-in-children [Accessed 3 December 2016].

PUBLIC HEALTH ENGLAND. 2015b. *Sugar reduction: the evidence for action.* Available: www.gov.uk/government/uploads/system/uploads/attachment_data/file/470179/Sugar_reduction_The_evidence_for_action.pdf. [Accessed 12 january 2016].

PUBLIC HEALTH ENGLAND. 2017. *Sugar Reduction: Achieving the 20%. A technical report outlining progress to date, guidelines for industry, 2015 baseline levels in key foods and next steps*. Available:https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/604336/Sugar_reduction_achieving_the_20_.pdf [Accessed 10 November 2013].

PYNE, V. & MACDONALD, I. A. 2016. Update on carbohydrates and health: the relevance of the Scientific Advisory Committee on Nutrition report for children. *Arch Dis Child,* 101**,** 876-880.

QADRI, G., ALKILZY, M., FENG, Y. S. & SPLIETH, C. 2015. Overweight and dental caries: the association among German children. *International Journal of Paediatric Dentistry,* 2**5,** 174-182.

QUAIL, A., WILLIAMS, J., MCCRORY, C., MURRAY, A. & THORNTON, M. 2011. Sample design and response in wave 1 of the infant cohort (at 9 months) of Growing Up in Ireland. Dublin, Ireland: Department of Health and Children.

QUINLAN, J. R. 1993. C4. 5: Programs for empirical learning. San Francisco, CA: Morgan Kaufmann.

QUINONEZ, R., SANTOS, R., WILSON, S. & CROSS, H. 2001a. The relationship between child temperament and early childhood caries. *Pediatr Dent,* 23**,** 5-10.

QUINONEZ, R. B., KEELS, M., VANN JR, W., MCIVER, F., HELLER, K. & WHITT, J. 2001b. Early Childhood Caries: analysis of psychosocial and

biological factors in a high–risk population. *Caries research,* 35**,** 376-383.

RADLOFF, L. S. 1977. The CES-D scale a self-report depression scale for research in the general population. *Applied psychological measurement,* 1**,** 385-401.

RAHMAN, M. M. & DAVIS, D. 2013. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing,* 3**,** 224-228.

REISINE, S., LITT, M. & TINANOFF, N. 1994. A biopsychosocial model to predict caries in preschool children. *Pediatr Dent,* 16**,** 413-418.

RENZAHO, A. & SILVA-SANIGORSKI, A. 2013. The importance of family functioning, mental health and social and emotional well-being on child oral health. *Child Care Health Dev,* 40**,** 543-552.

RIGGS, E., GIBBS, L., KILPATRICK, N., GUSSY, M., VAN GEMERT, C., ALI, S. & WATERS, E. 2015. Breaking down the barriers: a qualitative study to understand child oral health in refugee and migrant communities in Australia. *Ethnicity & health,* 20**,** 241-257.

ROBINSON, A. P., DUURSMA, R. A. & MARSHALL, J. D. 2005. A regression-based equivalence test for model validation: shifting the burden of proof. *Tree physiology,* 25**,** 903-913.

ROCKHILL, B. 2001. The privatization of risk. *American Journal of Public Health,* 91**,** 365.

ROGERS, R. D. 1985. The western information society. *New Information Technologies and Libraries.* Springer, 11-18.

ROKACH, L. & MAIMON, O. 2009. Classification trees. *Data mining and knowledge discovery handbook.* Springer, 149-174.

ROLLAND-CACHERA, M. F. 2011. Childhood obesity: current definitions and recommendations for their use. *Int J Pediatr Obes,* 6**,** 325-31.

ROSE, G. 1992. The strategy of preventive medicine. *The strategy of preventive medicine.*

ROSVALL, M. & BERGSTROM, C. T. 2010. Mapping change in large networks. *PLoS One,* 5**,** e8694.

ROTHMAN, K. J. 2017. The growing rift between epidemiologists and their data. *Eur J Epidemiol,* 32**,** 863-865.

ROTHMAN, K. J., GREENLAND, S., POOLE, C. & LASH, T. L. 2008. Causation and Causal Inference. *In:* ROTHMAN, K. J., GREENLAND, S. & LASH, T. L. (eds.) *Modern epidemiology.* 3rd ed.: Lippincott Williams & Wilkins, 5-31.

RUIZ, E., RODRIGUEZ, P., VALERO, T., ÁVILA, J., ARANCETA-BARTRINA, J., GIL, Á., GONZÁLEZ-GROSS, M., ORTEGA, R., SERRA-MAJEM, L.

& VARELA-MOREIRAS, G. 2017. Dietary Intake of Individual (Free and Intrinsic) Sugars and Food Sources in the Spanish Population: Findings from the ANIBES Study. *Nutrients,* 9**,** 275.

RUTISHAUSER, I. H. E. 2007. Dietary intake measurements. *Public Health Nutr,* 8.

SACN. 2015. *Carbohydrates and Health Report: SACN (Scientific Advisory Committee on Nutrition)* [Online]. London: TSO. Available: https://www.gov.uk/government/publications/sacn-carbohydrates-and-health-report [Accessed 3 June 2016].

SAGHERI, D., MCLOUGHLIN, J. & NUNN, J. H. 2013. Dental caries experience and barriers to care in young children with disabilities in Ireland. *Quintessence Int,* 44**,** 159-69.

SALLIS, J. F., TAYLOR, W. C., DOWDA, M., FREEDSON, P. S. & PATE, R. R. 2002. Correlates of vigorous physical activity for children in grades 1 through 12: comparing parent-reported and objectively measured physical activity. *Pediatric Exercise Science,* 14**,** 30.

SATIJA, A., YU, E., WILLETT, W. C. & HU, F. B. 2015. Understanding nutritional epidemiology and its role in policy. *Adv Nutr,* 6**,** 5-18.

SCHENKER, N. & RAGHUNATHAN, T. E. 2007. Combining information from multiple surveys to enhance estimation of measures of health. *Stat Med,* 26**,** 1802-11.

SELWITZ, R. H., ISMAIL, A. I. & PITTS, N. B. 2007. Dental caries. *The Lancet,* 369**,** 51-59.

SHEIHAM, A. 2005. Oral health, general health and quality of life. *Bulletin of the World Health Organization,* 83**,** 644-644.

SHEIHAM, A. 2006. Dental caries affects body weight, growth and quality of life in pre-school children. *Br Dent J,* 201**,** 625-6.

SHEIHAM, A. 2007. Dietary effects on dental diseases. *Public Health Nutr,* 4.

SHEIHAM, A. & JAMES, W. P. 2014. A new understanding of the relationship between sugars, dental caries and fluoride use: implications for limits on sugars consumption. *Public Health Nutr,* 17**,** 2176-84.

SHEIHAM, A. & JAMES, W. P. 2015. Diet and Dental Caries: The Pivotal Role of Free Sugars Reemphasized. *J Dent Res,* 94**,** 1341-7.

SHEIHAM, A. & WATT, R. G. 2000. The common risk factor approach: a rational basis for promoting oral health. *Community Dentistry and Oral Epidemiology,* 28**,** 399-406.

SHELLIS, R., BARBOUR, M., JESANI, A. & LUSSI, A. 2013. Effects of buffering properties and undissociated acid concentration on dissolution of dental enamel in relation to pH and acid type. *Caries research,* 47**,** 601-611.

SHIM, J.-S., OH, K. & KIM, H. C. 2014. Dietary assessment methods in epidemiologic studies. *Epidemiology and health,* 36.

SHRIVASTAVA, A., MURRIN, C. & KELLEHER, C. C. 2014. Preschoolers' parent-rated health disparities are strongly associated with measures of adiposity in the Lifeways cohort study children. *BMJ Open,* 4**,** e005328.

SLACK-SMITH, L. 2012. How population-level data linkage might impact on dental research. *Community Dent Oral Epidemiol,* 40 Suppl 2**,** 90-4.

SLACK-SMITH, L., COLVIN, L., LEONARD, H., KILPATRICK, N., BOWER, C. & MESSER, L. B. 2009. Factors associated with dental admissions for children aged under 5 years in Western Australia. *Arch Dis Child,* 94**,** 517-523.

SLACK-SMITH, L. 2003. Dental visits by Australian preschool children. *J Paediatr Child Health,* 39**,** 442-445.

SLADE, G. D. 2001. Epidemiology of dental pain and dental caries among children and adolescents. *Community Dent Health,* 18**,** 219-27.

SLUIK, D., VAN LEE, L., ENGELEN, A. I. & FESKENS, E. J. 2016. Total, Free, and Added Sugar Consumption and Adherence to Guidelines: The Dutch National Food Consumption Survey 2007-2010. *Nutrients,* 8**,** 70.

SMITH, L., KATZ, L., EMERY, H., SIEPPERT, J., POLSKY, Z. & NAGAN, K. 2014. It's about more than just baby teeth: An examination of early oral care in Canada. *Universal Journal of Public Health,* 2**,** 125-130.

SNIEHOTTA, F. F., ARAÚJO-SOARES, V., BROWN, J., KELLY, M. P., MICHIE, S. & WEST, R. 2017. Complex systems and individual-level approaches to population health: a false dichotomy? *The Lancet Public Health,* 2**,** e396-e397.

SOHN, W., BURT, B. A. & SOWERS, M. R. 2006. Carbonated soft drinks and dental caries in the primary dentition. *Journal of dental research,* 85**,** 262-266.

SONG, Y.-Y. & YING, L. 2015. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry,* 27**,** 130.

SPENCER, A. J. 2012. Putting the population back into oral health; decoupling oral health improvement from clinical dental practice. *Community Dent Oral Epidemiol,* 40 Suppl 2**,** 5-11.

SPITZ, A. S., WEBER-GASPARONI, K., KANELLIS, M. J. & QIAN, F. 2006. Child temperament and risk factors for early childhood caries. *Journal of dentistry for children,* 73.

STANHOPE, K. L. 2015. Sugar consumption, metabolic disease and obesity: The state of the controversy. *Critical reviews in clinical laboratory sciences***,** 1-16.

STATISTA. 2016. Available: https://www.statista.com/statistics/535219/global-sugar-per-capita-consumption-by-country/ [Accessed 11 February 2018].

STEPHEN, A., ALLES, M., DE GRAAF, C., FLEITH, M., HADJILUCAS, E., ISAACS, E., MAFFEIS, C., ZEINSTRA, G., MATTHYS, C. & GIL, A. 2012. The role and requirements of digestible dietary carbohydrates in infants and toddlers. *Eur J Clin Nutr,* 66**,** 765-79.

SUN, Y., KAMEL, M. S., WONG, A. K. C. & WANG, Y. 2007. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition,* 40**,** 3358-3378.

SUN, Y., WONG, A. K. & KAMEL, M. S. 2009. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence,* 23**,** 687-719.

SVENSSON, A., LARSSON, C., EIBEN, G., LANFER, A., PALA, V., HEBESTREIT, A., HUYBRECHTS, I., FERNANDEZ-ALVIRA, J. M., RUSSO, P., KONI, A. C., DE HENAUW, S., VEIDEBAUM, T., MOLNAR, D., LISSNER, L. & CONSORTIUM, I. 2014. European children's sugar intake on weekdays versus weekends: the IDEFICS study. *Eur J Clin Nutr,* 68**,** 822-8.

TALEKAR, B. S., ROZIER, R. G., SLADE, G. D. & ENNETT, S. T. 2005. Parental perceptions of their preschool-aged children's oral health. *The Journal of the American Dental Association,* 136**,** 364-372.

TANG, C., QUINONEZ, R. B., HALLETT, K., LEE, J. Y. & KENNETH WHITT, J. 2005. Examining the association between parenting stress and the development of early childhood caries. *Community Dentistry and Oral Epidemiology,* 33**,** 454-460.

TAREN, D., DWYER, J., FREEDMAN, L. & SOLOMONS, N. W. 2006. Dietary assessment methods: where do we go from here? *Public Health Nutr,* 5**,** 1001-1003.

TE MORENGA, L., MALLARD, S. & MANN, J. 2013. Dietary sugars and body weight: systematic review and meta-analyses of randomised controlled trials and cohort studies. *BMJ,* 346**,** e7492.

THAM, R., BOWATTE, G., DHARMAGE, S., TAN, D., LAU, M., DAI, X., ALLEN, K. & LODGE, C. 2015. Breastfeeding and the risk of dental caries: a systematic review and meta-analysis. *Acta paediatrica,* 104**,** 62-84.

THOMPSON, F. & SUBAR, A. 2013. Dietary assessment methodology. *In:* COULSTON, A., BOUSHEY, C. & FERRUZZI, M. (eds.) *Nutrition in the Prevention and Treatment of Disease.* 3rd ed. San Diego, CA: Academic Press, 1-46.

THOMPSON, F. E. & SUBAR, A. F. 2008. Dietary assessment methodology. *Nutrition in the Prevention and Treatment of Disease,* 2**,** 3-39.

TICKLE, M. & MILSOM, K. 2008. The whole population approach to caries prevention in general dental practice. *Br Dent J,* 205**,** 521.

TINANOFF, N., KANELLIS, M. & VARGAS, C. 2002. Current understanding of the epidemiology, mechanisms, and prevention of dental caries in preschool children. *Pediatr Dent,* 24**,** 543-551.

TINANOFF, N. & PALMER, C. A. 2000. Dietary determinants of dental caries and dietary recommendations for preschool children. *J Public Health Dent,* 60**,** 197-206.

TINANOFF, N. & REISINE, S. 2009. Update on early childhood caries since the Surgeon General's Report. *Academic pediatrics,* 9**,** 396-403.

TOOZE, J. A., KIPNIS, V., BUCKMAN, D. W., CARROLL, R. J., FREEDMAN, L. S., GUENTHER, P. M., KREBS-SMITH, S. M., SUBAR, A. F. & DODD, K. W. 2010. A mixed-effects model approach for estimating the distribution of usual intake of nutrients: the NCI method. *Stat Med,* 29**,** 2857-68.

TRAMINI, P., MOLINARI, N., TENTSCHER, M., DEMATTEI, C. & SCHULTE, A. G. 2009. Association between caries experience and body mass index in 12-year-old French children. *Caries Res,* 43**,** 468-73.

TUFTE, E. 2001. *The visual display of quantitative information,* Cheshire, USA, Graphics Press.

US Department of Health and Human Services, 2000. *Oral health in America: a report of the surgeon general—executive summary.* U.S. Public Health Service: Washington D.C., U.S.A.

VAN DER TAS, J. T., KRAGT, L., VEERKAMP, J. J., JADDOE, V. W., MOLL, H. A., ONGKOSUWITO, E. M., ELFRINK, M. E. & WOLVIUS, E. B. 2016. Ethnic Disparities in Dental Caries among Six-Year-Old Children in the Netherlands. *Caries research,* 50**,** 489-497.

VAN NOORDEN, R. 2015. Interdisciplinary research by the numbers. *Nature News,* 525**,** 306.

VANIA, A., PARISELLA, V., CAPASSO, F., DI TANNA, G. L., VESTRI, A., FERRARI, M. & POLIMENI, A. 2011. Early childhood caries underweight or overweight, that is the question. *European Journal of Paediatric Dentistry,* 12**,** 231.

VARGAS, C. M. & RONZIO, C. R. 2006. Disparities in early childhood caries. *BMC Oral Health,* 6 Suppl 1**,** S3.

WAGNER, Y. & HEINRICH-WELTZIEN, R. 2017. Risk factors for dental problems: Recommendations for oral health in infancy. *Early Hum Dev,* 114**,** 16-21.

WAKE, M., CLIFFORD, S., YORK, E., MENSAH, F., GOLD, L., BURGNER, D. & DAVIES, S. 2014. Introducing growing up in Australia's child health

check point: A physical health and biomarkers module for the longitudinal study of Australian children. *Family Matters*, 15.

WAKE, M., HARDY, P., CANTERFORD, L., SAWYER, M. & CARLIN, J. 2007. Overweight, obesity and girth of Australian preschoolers: prevalence and socio-economic correlates. *International Journal of Obesity,* 31, 1044-1051.

WAKE, M., HARDY, P., SAWYER, M. G. & CARLIN, J. B. 2008. Comorbidities of overweight/obesity in Australian preschoolers: a cross-sectional population study. *Arch Dis Child,* 93, 502-7.

WALTON, J. 2012. National Pre-School Nutrition Survey. *Summary report on: Food and Nutrient Intakes, Physical Measurements and Barriers to Healthy Eating*. Irish Universities Nutrition Alliance.

WALTON, J., EVANS, K., KEHOE, L., MCNULTY, B., NUGENT, A. & FLYNN, A. 2016. Intakes and sources of dietary sugars in Irish pre-school children aged 1-4 years. *Proc Nutr Soc,* 75.

WALTON, J. & FLYNN, A. 2013. Nutritional adequacy of diets containing growing up milks or unfortified cow's milk in Irish children (aged 12-24 months). *Food Nutr Res,* 57.

WALTON, J., KEHOE, L., MCNULTY, B. A., NUGENT, A. P. & FLYNN, A. 2017. Nutrient intakes and compliance with nutrient recommendations in children aged 1-4 years in Ireland. *J Hum Nutr Diet,* 30, 665-676.

WATT, R. G. 2007. From victim blaming to upstream action: tackling the social determinants of oral health inequalities. *Community Dent Oral Epidemiol,* 35, 1-11.

WATT, R. G. & ROUXEL, P. L. 2012. Dental caries, sugars and food policy. *Arch Dis Child,* 97, 769-772.

WATT, R. G. & SHEIHAM, A. 2012. Integrating the common risk factor approach into a social determinants framework. *Community Dentistry and Oral Epidemiology,* 40, 289-296.

WEGMAN, E. J. 2003. Visual data mining. *Stat Med,* 22, 1383-97.

WELLEK, S. 2010. *Testing statistical hypotheses of equivalence and noninferiority,* Florida, CRC Press.

WELSH, J. A. & FIGUEROA, J. 2017. Intake of Added Sugars During the Early Toddler Period. *Nutr Today,* 52, S60-S68.

WELSH, J. A., SHARMA, A. J., GRELLINGER, L. & VOS, M. B. 2011. Consumption of added sugars is decreasing in the United States. *Am J Clin Nutr,* 94, 726-34.

WHELTON, H., O'MULLANE, D., HARDING, M., GUINEY, H., CRONIN, M., FLANNERY, E. & KELLEHER, V. 2006. North South survey of

children's oral health in Ireland 2002. Dublin, Ireland: Brunswick Press Ltd.

WICKHAM, H. 2010. A Layered Grammar of Graphics. *Journal of Computational and Graphical Statistics,* 19**,** 3-28.

WICKHAM, H. & GROLEMUND, G. 2016*. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Available: http://r4ds.had.co.nz/ [Accessed 3 January 2017].

WILLIAMS, G. 2011. *Data mining with Rattle and R: The art of excavating data for knowledge discovery*, Springer Science & Business Media.

WILLIAMS, J., MURRAY, A., MCCRORY, C. & MCNALLY, S. 2013. Development from birth to three years - Growing Up in Ireland. Dublin, Ireland: Office of the Minister for Children and Youth Affairs.

WITTEKIND, A. & WALTON, J. 2014. Worldwide trends in dietary sugars intake. *Nutr Res Rev,* 27**,** 330-45.

WOODWARD, M. 2013. *Epidemiology: study design and data analysis,* Florida, CRC press.

WORLD HEALTH ORGANIZATION 2000. *Obesity - Preventing & Managing the Global Epidemic*. WHO technical report series*.* Geneva: WHO.

WORLD HEALTH ORGANIZATION 2002. *Globalization, diets and noncommunicable diseases*. Geneva: WHO.

WORLD HEALTH ORGANIZATION 2003. *Diet, nutrition and the prevention of chronic diseases*. WHO technical report series*.* Geneva: WHO.

WORLD HEALTH ORGANIZATION 2015*. Guideline: sugars intake for adults and children*. Geneva: WHO.

WORLD HEALTH ORGANIZATION 2017a. *Incentives and disincentives for reducing sugar in manufactured foods. An exploratory supply chain analysis. A set of insights for Member States in the context of the WHO European Food and Nutrition Action Plan 2015–2020*. Copenhagen, Denmark: WHO.

WORLD HEALTH ORGANIZATION. 2017b. *WHO expert consultation on public health intervention against early childhood caries: report of a meeting, Bangkok, Thailand, 26-28 January 2016* [Online]. Geneva: WHO. Available: http://apps.who.int/iris/handle/10665/255627 [Accessed 11 October 2017].

YAP, B. W., RANI, K. A., RAHMAN, H. A. A., FONG, S., KHAIRUDIN, Z. & ABDULLAH, N. N. 2014. An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013),* 285**,** 13-22.

YOE, C. 2011. *Principles of risk analysis: decision making under uncertainty,* New York, CRC press.

YOO, C., RAMIREZ, L. & LIUZZI, J. 2014. Big data analysis using modern statistical and machine learning methods in medicine. *Int Neurourol J,* 18**,** 50-7.

YOO, I., ALAFAIREET, P., MARINOV, M., PENA-HERNANDEZ, K., GOPIDI, R., CHANG, J. F. & HUA, L. 2012. Data mining in healthcare and biomedicine: a survey of the literature. *J Med Syst,* 36**,** 2431-2448.

ZERO, D. 2004. Sugars–the arch criminal? *Caries research,* 38**,** 277-285.

ZUMEL, N., MOUNT, J. & PORZAK, J. 2014. *Practical data science with R,* New York, Manning.

Appendix A

Table A1 Index to selected code fragments and code documents.

Chapter	Data	Details	Code fragments and/or code file names	page
3, 4	GUI_Wave-1 GUI_Wave-2	Decision trees	SPSS code fragments	217 218
4	GUI_Wave-2	Calculation of IOTF cut-offs for reclassification of BMI Calculation of lms for UK/WHO obesity cut-offs	Code fragments	219 220
5	NPNS GUI_Wave-2	Data linkage, mapping, aligning, filter, equivalence tests, proportion tests	IUNA- NPNS- mapped to GUI_Rmd	224
5,6		Data linkage, non-covered analysis	Non-covered_Analysis.RMd	227
6	NPNS GUI_Wave-2	Cariogenic food meal analysis, Association analysis	Cariogenic_summary.Rmd Association analysis.Rmd	249 280
7	NPNS GUI_Wave-2	Free sugar mapping comparison Analysis of total, free sugar distributions covered by GUI and NPNS Recategorisation of NPNS food codes to assess key food sources	Free-sugars-mapping.Rmd Non-covered-Sugars.Rmd Uncovered IUNA Food code mapping.csv	303 313 338

Subsection 1. SPSS code fragments: Chapters 3 and 4

Note: All DataSet files used were restricted access, RMF files from the GUI infant cohort, Wave 1 or Wave 2(access only via Central Statistics Office application). Key Code fragments are highlighted in yellow, e.g., misclassification costs.

Chapter 3

a. SPSS code for classification tree (Wave 1, 9 Months old)

```
DATASET ACTIVATE DataSet RMF file GUI-Wave 1

TREE apch38f_Rec [n] BY aphc01b [s] aphc01a [n] aphc02a
[n] apch40 [n] aded07a [s] aded07b [s]

    aded07c [s] aded07d [s] adfc18p [s] adfc19p [s]
adph24p [s] adph25p [s] apsd43a [n] apsd53 [n]

    adsd56a [s] adsd58a [s] apcl15 [n] apch01bcat [n]
apph02_Rec_Binary [n] apcl07 [n]

  /TREE         DISPLAY=TOPDOWN         NODES=STATISTICS
BRANCHSTATISTICS=YES NODEDEFS=YES SCALE=AUTO

  /DEPCATEGORIES USEVALUES= [.00 1.00] TARGET= [1.00]

  /PRINT MODELSUMMARY CLASSIFICATION RISK

  /GAIN  CATEGORYTABLE=YES  TYPE=[NODE]  SORT=DESCENDING
CUMULATIVE=NO

  /METHOD TYPE=CHAID

  /GROWTHLIMIT       MAXDEPTH=5       MINPARENTSIZE=100
MINCHILDSIZE=50

  /VALIDATION      TYPE=      CROSSVALIDATION      (10)
OUTPUT=BOTHSAMPLES

  /CHAID ALPHASPLIT=0.05 ALPHAMERGE=0.05 SPLITMERGED=NO
CHISQUARE=PEARSON CONVERGE=0.001

    MAXITERATIONS=100 ADJUST=BONFERRONI INTERVALS=10

  /COSTS CUSTOM= .00 .00 [0] .00 1.00 [.05] 1.00 .00 [.95]
1.00 1.00 [0]
```

b. SPSS code for classification tree (3 years of age)

```
DATASET ACTIVATE DataSet RMF file GUI-Wave 2

TREE bpch57 [n] BY bphc02a [n] bphc01a [n] bpph02_Rec_Bin
[n] bpch01b_Rec_Bin [n] bpch47a [n]

    bpch56 [n] bdcr04p [n] bdcr05p [s] bpcl17a [n]
bpcl17b [n] bded19a [s] bded19b [s] bded19c [s]

    bded20f [s] bdfc21p [s] bpph21 [n] bdph24p [s]
bdph25p [n] bpsd43a [n] bpsd53 [n] bdsd56a [n]

    bdsd58a [s]

  /TREE         DISPLAY=TOPDOWN         NODES=STATISTICS
BRANCHSTATISTICS=YES NODEDEFS=YES SCALE=AUTO

  /DEPCATEGORIES USEVALUES= [1 2] TARGET= [1 2]

  /PRINT MODELSUMMARY CLASSIFICATION RISK

  /GAIN  CATEGORYTABLE=YES  TYPE=[NODE]  SORT=DESCENDING
CUMULATIVE=NO

  /METHOD TYPE=CHAID

  /GROWTHLIMIT       MAXDEPTH=4       MINPARENTSIZE=100
MINCHILDSIZE=50

  /VALIDATION         TYPE=CROSSVALIDATION         (10)
OUTPUT=BOTHSAMPLES

  /CHAID ALPHASPLIT=0.05 ALPHAMERGE=0.05 SPLITMERGED=NO
CHISQUARE=PEARSON CONVERGE=0.001

    MAXITERATIONS=100 ADJUST=BONFERRONI INTERVALS=10

  /COSTS CUSTOM= 1 1 [0] 1 2 [.95] 2 1 [.05] 2 2 [0]

 /MISSING NOMINALMISSING=MISSING
```

Chapter 4

a. Recoding of BMI classification for RMF datafiles for 3 year olds using the IOTF classification

```
Label: RMF data recoded BMI according to IOTF
classification

GET

  FILE=DATASET ACTIVATE DataSet RMF file GUI-Wave 2

*recode BMI to male female IOTF.

DO IF (bphc02a = 1).

RECODE bdpm06c (MISSING=0) (Lowest thru 14.82999=1)
(14.83 thru 17.84999=2) (17.85 thru

    19.49999=3)    (19.5    thru    Highest=4)    INTO
BMI_Class_IOTF_Male.

END IF.

EXECUTE.

DO IF (bphc02a = 2).

RECODE bdpm06c (MISSING=0) (14.60 thru 17.63999=2)
(Lowest thru 14.599999=1) (17.64 thru

    19.37999=3)    (19.38    thru    Highest=4)    INTO
BMI_Class_IOTF_Female.

END IF.

EXECUTE.

FREQUENCIES VARIABLES=BMI_Class_IOTF_Female

  /ORDER=ANALYSIS.

COMPUTE       BMI_class_IOTF=BMI_Class_IOTF_Male       +
BMI_Class_IOTF_Female.

EXECUTE.
```

b. LMS method for BMI classification of 3 year olds using WHO/UK cut-offs

```
DATASET ACTIVATE DataSet RMF file GUI-Wave 2 infant
cohort

WEIGHT BY bzwg01.

FREQUENCIES VARIABLES=BMI_IOTF_classification_3YO

  /ORDER=ANALYSIS.

IF (bphc02a = 1) z_ind_male=((bdpm06c / 15.5988)  ** (-
0.3101) - 1)/ (-0.3101*0.07931).

EXECUTE.

IF (bphc02a = 2) z_ind_female=((bdpm06c / 15.3968)  ** (
     -0.5684) - 1)/ (-0.5684*0.08535).

EXECUTE.

IF           (ABS(z_ind_male)           <=           3)
z_ind_male_abs_lt_3=z_ind_male.

EXECUTE.

IF (z_ind_male > 3) z_ind_male_gt3= 3+ (bdpm06c - 20.0)
/ (20.0 -18.4).

EXECUTE.

IF (z_ind_male < - 3) z_ind_male_lt_3= -3 + (bdpm06c -
12.4 ).

EXECUTE.

IF           (ABS(z_ind_female)           <=           3)
z_ind_female_abs_lt_3=z_ind_female.

EXECUTE.
```

```
IF (z_ind_female > 3) z_ind_female_gt3= 3+ (bdpm06c -
20.3) / (20.3 -18.4).

EXECUTE.

IF (z_ind_female < - 3) z_ind_female_lt_3= -3 + (bdpm06c
- 12.1 ).

EXECUTE.

COMPUTE      z_final_male=SUM      (z_ind_male_abs_lt_3,
z_ind_male_gt3, z_ind_male_lt_3).

EXECUTE.

COMPUTE z_final_female = SUM ( z_ind_female_abs_lt_3,
z_ind_female_gt3,  z_ind_female_lt_3).

EXECUTE.

COMPUTE
percentile_male=100*CDF.NORMAL(z_final_male,0,1).

EXECUTE.

COMPUTE
percentile_female=100*CDF.NORMAL(z_final_female,0,1).

EXECUTE.

COMPUTE      percentile      =      SUM(percentile_male,
percentile_female).

EXECUTE.

IF (percentile <= 91) bmi_level_0=0.

EXECUTE.

IF (percentile > 91 AND percentile <= 98) bmi_level_1=1.

EXECUTE.

IF (percentile > 98) bmi_level_2=2.
```

```
EXECUTE.

COMPUTE bmi_level = SUM (bmi_level_0, bmi_level_1,
bmi_level_2).

EXECUTE.

VALUE LABELS bmi_level

  .00 'Normal'

  1.00 'Overweight'

  2.00 'Obese'.

EXECUTE.
```

c. SPSS code for classification tree (3 years of age)

```
TREE bpch57 [n] BY bphc02a [n] bpch05b [n] bpch03 [n]
bpch47a [n] bpch56 [n] bpcn03a [n] bpcn03c

    [n] bpcn09a [n] bpcn09b [n] bpcn09c [n] bpcn09d [n]
bpcn09e [n] bpcn09f [n] bpcn09g [n] bpcn09h [n]

    bpcn09i [n] bpcn09j [n] bpcn09k [n] bpcn09l [n]
bpcn09m [n] bpcn09n [n] bpcn09o [n] bpcl17a [n]

    bpcl17b [n] bpcl18aa [s] bdsd58a [s] bdsd56a [n]
bpsd53 [n] bpsd43a [n] bdpm06p [s] BMI_class_IOTF

    [n]

  /TREE DISPLAY=TOPDOWN NODES=STATISTICS
BRANCHSTATISTICS=YES NODEDEFS=YES SCALE=AUTO

  /DEPCATEGORIES USEVALUES= [1 2] TARGET=[1]

  /PRINT MODELSUMMARY CLASSIFICATION RISK

  /GAIN CATEGORYTABLE=YES TYPE=[NODE] SORT=DESCENDING
CUMULATIVE=NO

  /METHOD TYPE=CHAID

  /GROWTHLIMIT MAXDEPTH=5 MINPARENTSIZE=100
MINCHILDSIZE=50

  /VALIDATION TYPE= CROSSVALIDATION (10)
OUTPUT=BOTHSAMPLES

  /CHAID ALPHASPLIT=0.05 ALPHAMERGE=0.05 SPLITMERGED=NO
CHISQUARE=PEARSON CONVERGE=0.001

    MAXITERATIONS=100 ADJUST=BONFERRONI INTERVALS=10

  /COSTS CUSTOM= 1 1 [0] 1 2 [.95] 2 1 [.05] 2 2 [0]

  /MISSING NOMINALMISSING=MISSING.
```

Subsection 2. Rmarkdown and RStudio files for Chapters 4-7.

1. IUNA NPNS mapped to GUI. Rmd

Key packages

- `SQLDF` to join data: joins data more effectively using SQL (link to http://www.iana.org/assignments/media-types/application/sql)
- `foreign` read the SPSS data

```
library(sqldf)

'foreign' was built under R version 3.4.1

storePath <- "tmp"
```

NPNS data was in two separate files

1. Antropometric data
2. Food diary

Joined both files as needed the antropometry data in the analysis

Join is on subject id

```
# Read the two datasets
antropoDf <- read.csv("npns-antropometry-data.csv", header = T
RUE)
fooddataDf <- read.csv("npns-derived-v1_copy.csv", header = TR
UE)

# Join the two dataset using the id of the subject
foodDataAgeDf <-
sqldf("SELECT f.*, a.* FROM fooddataDf f LEFT JOIN antropoDf a
ON f.SUBJECID = a.ID")
```

Limited to 3 years old

```
# Looking at only 3 and 5 years old - Note there is no 5 years
subjects
foodDataForMapping <- sqldf("SELECT * from foodDataAgeDf WHERE
AGE = 3")
```

The GUI food questionnaire and the NPNS data in SPSS format

```
# Reading in the GUI codes
guiFoodQuestions <- read.table(file="GUI-qn.txt", sep=".")
colnames(guiFoodQuestions) <- c('code', 'description')
foodDataSPSS <- read.spss(file="npns-food-file-4R.sav")

## re-encoding from CP1252
```

Included the cooking methods description to build the dataset that is used for mapping using the following variables

3. Food code
4. Food description (short and long)
5. Cooking method (code and description)
6. F77 food category

```
# Loaded initially to map against the 77 food categories
iunaFoodCodes <-
cbind.data.frame(code = seq(1, 77),
description = levels(foodDataSPSS$IUNA_NPNS_77FG))

# Loaded initially to use the cooking
cookingMethodsCodes <- read.table(file = "CMETH.txt", sep = "=
")

colnames(cookingMethodsCodes) <- c('code', 'description')

# Group diary entries by IUNA Code 77 and cooking method.
foodGroupsCmeth <-
sqldf("SELECT IUNA_NPNS_77FG, CMETH, count(*) ct FROM foodData
ForMapping GROUP BY 1,2 ")

# Add description to the cooking method
foodGroupsCmethExt <-
sqldf(
"SELECT fc.description FOODNAME, cm.description COOKINGMETHOD
, fg.*
FROM foodGroupsCmeth fg
LEFT JOIN cookingMethodsCodes cm ON cm.code = fg.CMETH
LEFT JOIN iunaFoodCodes fc ON fc.code = fg.IUNA_NPNS_77FG"
)


write.csv(
foodGroupsCmethExt,
file = paste(storePath, "iuna-food-groups-cooking-methods.csv"
, sep = "/"),
row.names = FALSE
)
```

```
# List of all distinct food in the diary
allFoods <-
sqldf(
"SELECT distinct IUNA_NPNS_77FG, CMETH, FCODE, Food_descriptio
n_first_first
FROM fooddataDf  "
)
## IUNA_NPNS_77FG is too coarse we need to use the actual food
codes


allFoodsExt <-
sqldf(
"SELECT fc.description FOODNAME, cm.description COOKINGMETHOD
, fg.*
FROM allFoods fg
LEFT JOIN cookingMethodsCodes cm ON cm.code = fg.CMETH
LEFT JOIN iunaFoodCodes fc ON fc.code = fg.IUNA_NPNS_77FG"
)

write.csv(
allFoodsExt,
file = paste(storePath, "iuna-food-allFoods-with-groups.csv",
sep = "/"),
row.names = FALSE
)
```

Mapping done in Google sheet with filters

At this point the mapping was read. If not mapped we set the `GUI_CODE` to `NULL`

```
mappings <- read.csv(file=paste("iuna-gui-mapping-2015-12-21.c
sv", sep="/"), header = TRUE)

# Setting all empty i.e non categorized foods to NA
mappings[mappings$GUI_CODE == "",]$GUI_CODE <- NA
```

Joined the mapping back to the NPNS data using:

7. FOOD CODE
8. COOKING METHOD

```
foodDataGUIMapped <-
  sqldf(
  "SELECT f.*, m.GUI_CODE  from foodDataForMapping f
  LEFT JOIN mappings m on m.FCODE = f.FCODE AND m.CMETH = f.CM
ETH"
  )
```

2. Non-covered analysis.Rmd.

Initially, the data was loaded after we mapped and the IUNA NPNS 77 Food Group was used to group the different foods.

The initial aggregation was done at `SUBJECT_ID` and `SURVDAY` survey day meaning that for each subject and each day of the survey we got an aggregated record.

Defined 4 aggregate metrics:

1. `non_gui_ct` number of entries in the diary which do not map to a GUI food code. Note that an entry in the diary is a *consumption*.
2. `non_gui_fwt` the total weight of the entries which do not map to a GUI food code.
3. `day_ct` total number of entries.
4. `day_fwt` total weight of the entries.

Note that all the 4 aggreates are agreggated at subject and survey day level, meaning that there is an entry for each subject and for each day of the survey.

```
foodDataGUIMapped <- read.csv("foodDataGUIMappedV2.csv")
foodDataSPSS <- read.spss(file="npns-food-file-4R.sav")

DOW <- c('SUN', 'MON', 'TUE', 'WED', 'THU', 'FRI', 'SAT')
foodDataGUIMapped$DOW <- factor(sapply(foodDataGUIMapped$DOW,
function(x){

    DOW[x]
}), levels = DOW)
# Loaded initially to map against the 77 food categories
iunaFoodCodes<- cbind.data.frame(code=seq(1,77), description=l
evels(foodDataSPSS$IUNA_NPNS_77FG))


nonGUIConsumptions <- sqldf("SELECT SUBJECID, DOW,
                            SUM(CASE WHEN GUI_CODE IS NULL THE
N 1 ELSE 0 END) non_gui_ct,
                            SUM(CASE WHEN GUI_CODE IS NULL THE
N FWT ELSE 0 END) non_gui_fwt,
                            SUM(CASE WHEN GUI_CODE IS NULL THE
N sugars ELSE 0 END) uncovered_total_sugar,
                            SUM(CASE WHEN GUI_CODE IS NULL THE
N 0 ELSE sugars END) covered_total_sugars,
                            SUM(sugars) total_sugars,
                            COUNT(*) day_ct,
                            SUM(FWT) day_fwt
```

```
                                FROM foodDataGUIMapped GROUP BY SU
BJECID, DOW")
```

Ratio of non-GUI and total consumptions

Checked if the ratio changes over the day of the survey. Assumed independance across the 4 days.

Overall view

```
library(ggplot2)
nonGUIConsumptions$ratio_ct <- nonGUIConsumptions$non_gui_ct/n
onGUIConsumptions$day_ct
nonGUIConsumptions$ratio_fwt <- nonGUIConsumptions$non_gui_fwt
/nonGUIConsumptions$day_fwt


par(mfrow=c(1,1))
xval <- nonGUIConsumptions$ratio_ct*100
h<-hist(xval, xlab='% of total count of consumptions per day',
     main="Unmapped consumptions ratio", col='orange')
xfit <- seq(min(xval), max(xval), length.out = 40)
yfit <- dnorm(xfit, mean = mean(xval), sd= sd(xval))
yfit <- yfit * diff(h$mids[1:2]) * length(xval)
lines(xfit, yfit, col='blue', lwd=2)
box()
```

```
xval <- nonGUIConsumptions$ratio_fwt*100
h<-hist(xval, xlab='% of total food weight per day',
     main="Unmapped food weight", col='orange')
xfit <- seq(min(xval), max(xval), length.out = 40)
yfit <- dnorm(xfit, mean = mean(xval), sd= sd(xval))
yfit <- yfit * diff(h$mids[1:2]) * length(xval)
lines(xfit, yfit, col='blue', lwd=2)
box()
```

```
## [1] "=====Uncovered Food Items Weight ====="

pastecs::stat.desc(nonGUIConsumptions[,c("non_gui_ct", "day_ct
", "non_gui_fwt", "day_fwt")])

##                 non_gui_ct       day_ct  non_gui_fwt      da
y_fwt
## nbr.val       504.0000000  504.0000000 5.040000e+02 5.04000
0e+02
```

```
## nbr.null         0.0000000    0.0000000 0.000000e+00 0.00000
0e+00
## nbr.na           0.0000000    0.0000000 0.000000e+00 0.00000
0e+00
## min              1.0000000    2.0000000 7.100000e+01 7.60000
0e+01
## max             21.0000000   35.0000000 1.307000e+03 2.46600
0e+03
## range           20.0000000   33.0000000 1.236000e+03 2.39000
0e+03
## sum           4057.0000000 9211.0000000 2.088480e+05 6.22553
0e+05
## median           8.0000000   18.0000000 3.900000e+02 1.21400
0e+03
## mean             8.0496032   18.2757937 4.143810e+02 1.23522
4e+03
## SE.mean          0.1349343    0.2223957 9.007997e+00 1.64196
4e+01
## CI.mean.0.95     0.2651043    0.4369389 1.769793e+01 3.22595
3e+01
## var              9.1764611   24.9277628 4.089658e+04 1.35880
7e+05
## std.dev          3.0292674    4.9927711 2.022290e+02 3.68620
1e+02
## coef.var         0.3763251    0.2731904 4.880268e-01 2.98423
6e-01
```

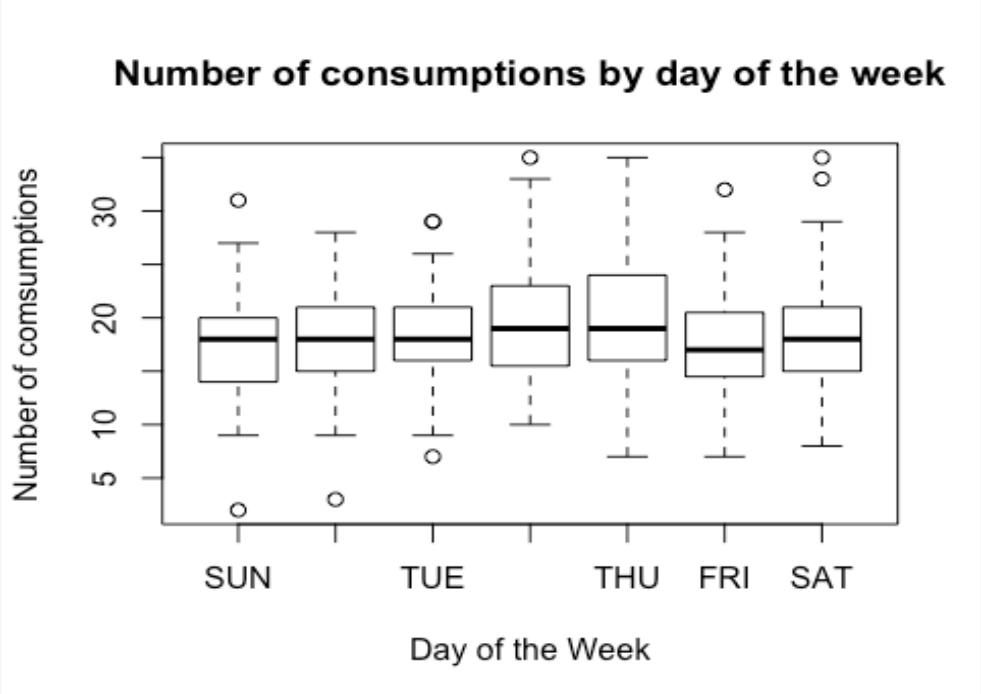
```
pastecs::stat.desc(nonGUIConsumptions[, c("ratio_ct", "ratio_f
wt")])
```

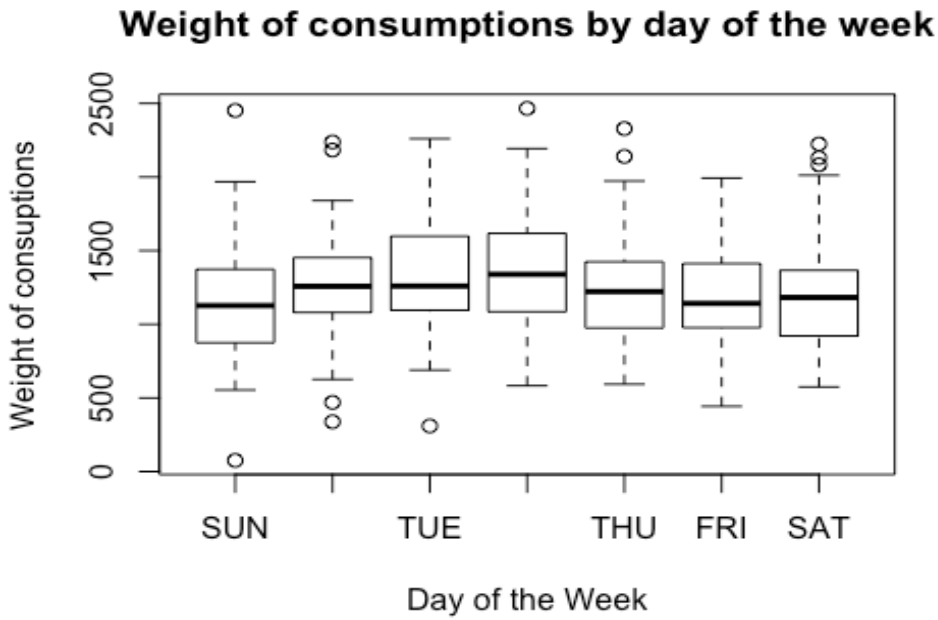
```
##                     ratio_ct    ratio_fwt
## nbr.val       5.040000e+02 5.040000e+02
## nbr.null      0.000000e+00 0.000000e+00
## nbr.na        0.000000e+00 0.000000e+00
## min           1.250000e-01 4.899931e-02
## max           1.000000e+00 1.000000e+00
## range         8.750000e-01 9.510007e-01
## sum           2.238548e+02 1.738821e+02
## median        4.444444e-01 3.323618e-01
## mean          4.441563e-01 3.450041e-01
## SE.mean       5.476348e-03 6.785794e-03
## CI.mean.0.95 1.075933e-02 1.333199e-02
## var           1.511516e-02 2.320769e-02
## std.dev       1.229437e-01 1.523407e-01
## coef.var      2.768028e-01 4.415620e-01
```

Analysed how the ratio of the covered food varies over each day of the week.

First looked at the distribution of the total number of consumptions and the total food over the day of the week.1 = Sunday; 2 = Monday; 3 = Tuesday; 4 = Wednesday; 5 = Thursday; 6 = Friday; 7 = Saturday

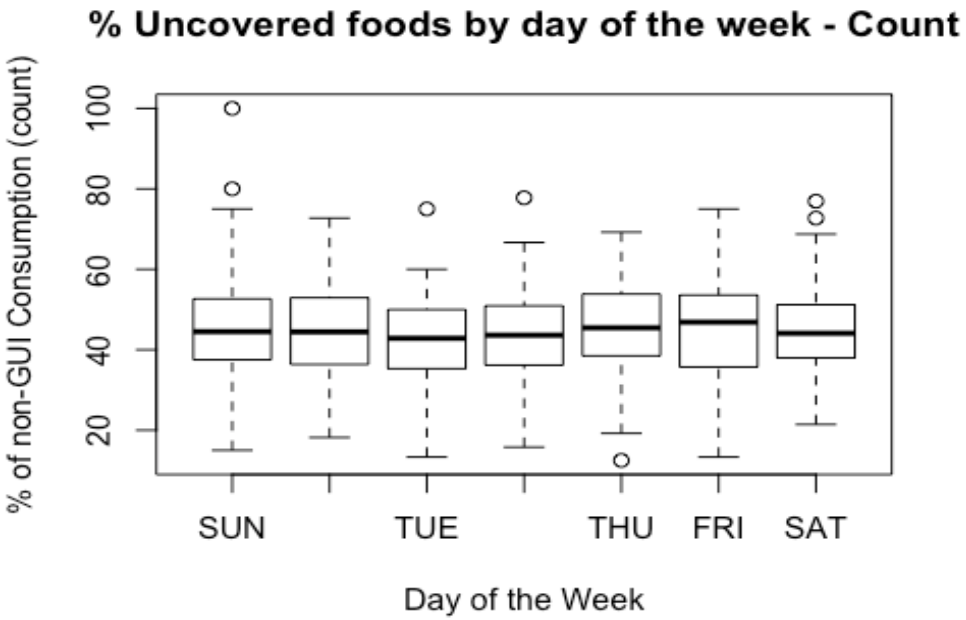
```
library(sm)

library(boot)
par(mfrow=c(1,1))

nonGUIConsumptions$dowFactored <- factor(nonGUIConsumptions$DOW)

with(nonGUIConsumptions,
     list(
     ct=boxplot(day_ct~DOW, xlab="Day of the Week", ylab="Number of comsumptions",
                main="Number of consumptions by day of the week - Count"),
     weight=boxplot(day_fwt~DOW, xlab="Day of the Week", ylab="Weight of consuptions",
                main="Weight of consuptions by day of the week - Weight")
```

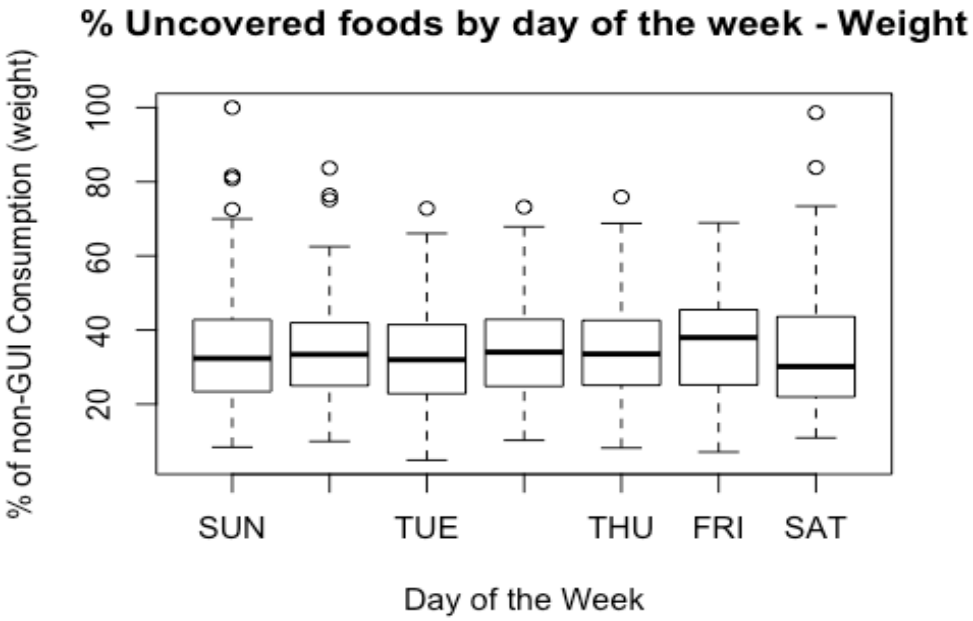




Then explored the actual ratio.

```
with(nonGUIConsumptions,
     list(
     ct=boxplot(ratio_ct*100~DOW, xlab="Day of the Week", ylab
="% of non-GUI Consumption (count)",
             main="% Uncovered foods by day of the week - Coun
t"),
     weight=boxplot(ratio_fwt*100~DOW, xlab="Day of the Week",
ylab="% of non-GUI Consumption (weight)",
             main="% Uncovered foods by day of the week - Weig
ht")
     )
)
```






```
reportedStats <- as.data.frame(aggregate(cbind(non_gui_ct, day
_ct, non_gui_fwt, day_fwt, ratio_ct, ratio_fwt)~DOW, data=nonG
UIConsumptions, function(x){
  rst <- pastecs::stat.desc(x)
  c(rst['nbr.val'], rst['min'], rst['max'], rst['range'], rst[
```

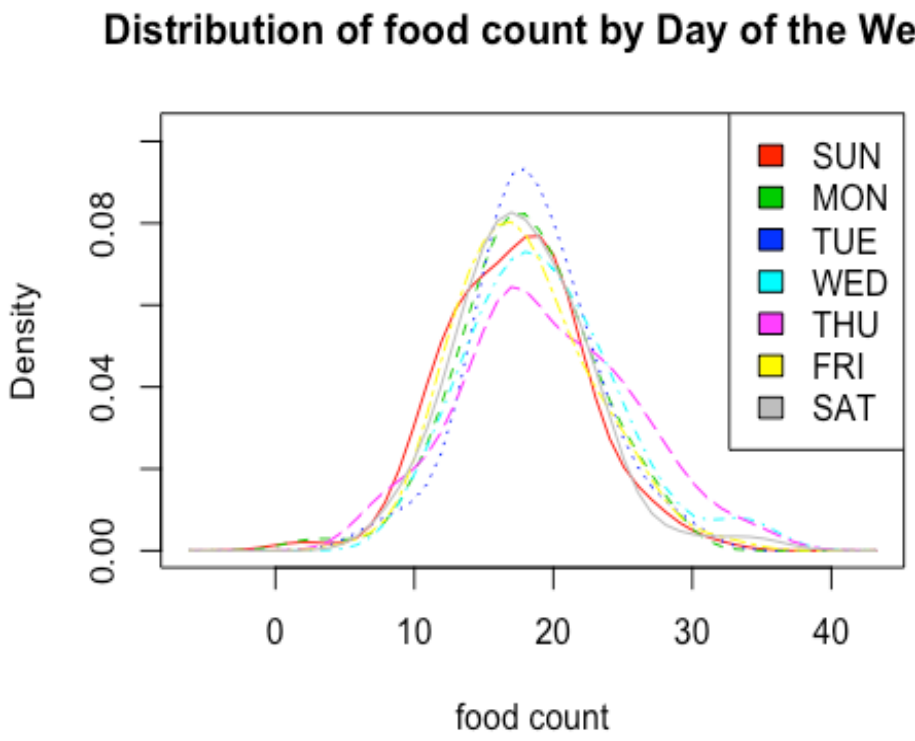
```
'median'], rst['mean'], rst['std.dev'])
} ))

kable(t(reportedStats[,-1 ]), col.names = reportedStats[,1 ],
digits = 1, description="descriptive for key variable")
```

	1	2	3	4	5	6	7
non_gui_ct.nbr.val	88.0	63.0	58.0	64.0	63.0	68.0	100.0
non_gui_ct.min	2.0	2.0	2.0	3.0	1.0	2.0	3.0
non_gui_ct.max	18.0	19.0	16.0	16.0	19.0	18.0	21.0
non_gui_ct.range	16.0	17.0	14.0	13.0	18.0	16.0	18.0
non_gui_ct.median	8.0	8.0	8.0	8.0	9.0	8.0	7.0
non_gui_ct.mean	7.7	8.0	7.8	8.3	8.8	8.0	7.9
non_gui_ct.std.dev	2.8	3.2	2.8	3.0	3.6	2.8	2.9
day_ct.nbr.val	88.0	63.0	58.0	64.0	63.0	68.0	100.0
day_ct.min	2.0	3.0	7.0	10.0	7.0	7.0	8.0
day_ct.max	31.0	28.0	29.0	35.0	35.0	32.0	35.0
day_ct.range	29.0	25.0	22.0	25.0	28.0	25.0	27.0
day_ct.median	18.0	18.0	18.0	19.0	19.0	17.0	18.0
day_ct.mean	17.3	18.0	18.5	19.4	19.6	17.9	17.9
day_ct.std.dev	4.8	4.7	4.3	5.4	6.2	4.7	4.7
non_gui_fwt.nbr.val	88.0	63.0	58.0	64.0	63.0	68.0	100.0
non_gui_fwt.min	76.0	78.0	71.0	138.0	100.0	96.0	83.0
non_gui_fwt.max	1144.0	1307.0	1016.0	1154.0	980.0	1240.0	1291.0
non_gui_fwt.range	1068.0	1229.0	945.0	1016.0	880.0	1144.0	1208.0
non_gui_fwt.median	351.0	445.0	393.5	411.5	406.0	393.0	360.0
non_gui_fwt.mean	383.9	436.9	422.4	453.8	419.7	414.4	393.9
non_gui_fwt.std.dev	203.0	224.7	199.2	208.6	167.9	206.9	200.4
day_fwt.nbr.val	88.0	63.0	58.0	64.0	63.0	68.0	100.0
day_fwt.min	76.0	338.0	310.0	583.0	593.0	443.0	575.0
day_fwt.max	2452.0	2238.0	2260.0	2466.0	2329.0	1993.0	2224.0
day_fwt.range	2376.0	1900.0	1950.0	1883.0	1736.0	1550.0	1649.0
day_fwt.median	1127.0	1258.0	1259.5	1339.5	1222.0	1142.0	1182.0
day_fwt.mean	1147.9	1263.6	1331.8	1351.3	1249.1	1178.8	1193.6
day_fwt.std.dev	372.0	345.3	380.0	389.5	372.1	330.8	358.0
ratio_ct.nbr.val	88.0	63.0	58.0	64.0	63.0	68.0	100.0
ratio_ct.min	0.2	0.2	0.1	0.2	0.1	0.1	0.2
ratio_ct.max	1.0	0.7	0.8	0.8	0.7	0.8	0.8
ratio_ct.range	0.8	0.5	0.6	0.6	0.6	0.6	0.6
ratio_ct.median	0.4	0.4	0.4	0.4	0.5	0.5	0.4
ratio_ct.mean	0.5	0.4	0.4	0.4	0.5	0.4	0.4
ratio_ct.std.dev	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ratio_fwt.nbr.val	88.0	63.0	58.0	64.0	63.0	68.0	100.0
ratio_fwt.min	0.1	0.1	0.0	0.1	0.1	0.1	0.1
ratio_fwt.max	1.0	0.8	0.7	0.7	0.8	0.7	1.0
ratio_fwt.range	0.9	0.7	0.7	0.6	0.7	0.6	0.9
ratio_fwt.median	0.3	0.3	0.3	0.3	0.3	0.4	0.3

ratio_fwt.mean	0.3	0.3	0.3	0.3	0.4	0.4	0.3
ratio_fwt.std.dev	0.2	0.2	0.1	0.1	0.1	0.1	0.2

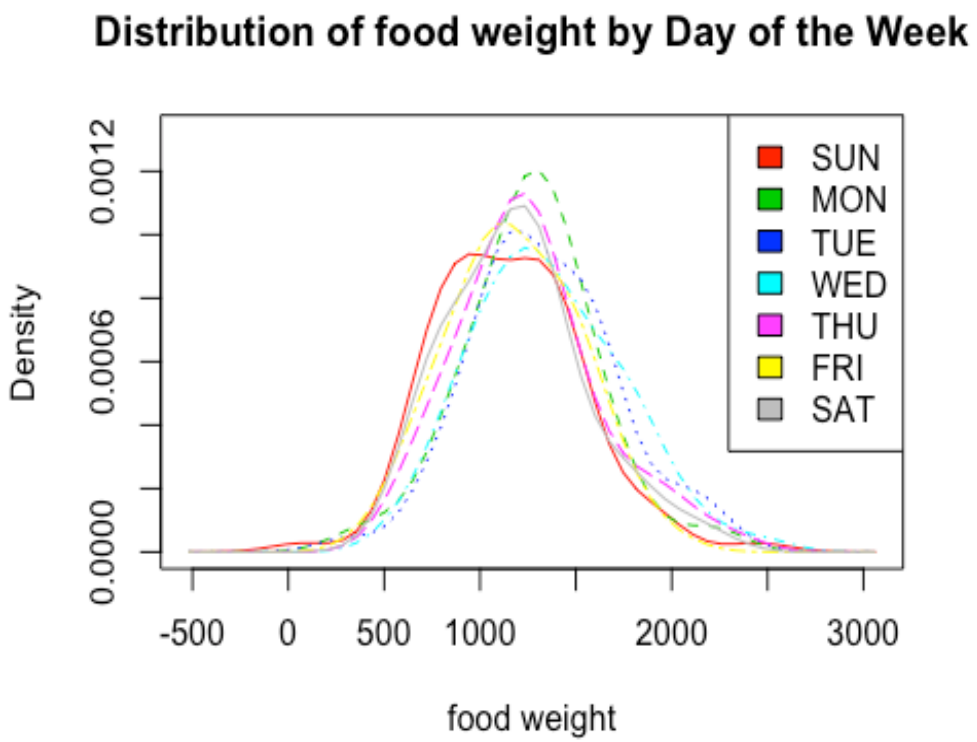
A nonparametric density estimation was used and the distribution of the 4 variables investigated: 1. `day_ct` total count of consumptions 2. `day_fwt` total weight of food in a day 3. `ratio_ct` the ratio of the number of consumptions that are covered by GUI 4. `ratio_fwt` the ratio of the food weight of the consumptions that are covered by GUI

```
colfill <- c(2:(1+length(nonGUIConsumptions$dowFactored)))

sm.density.compare(nonGUIConsumptions$day_ct, nonGUIConsumptio
ns$dowFactored, xlab="food count" )
title("Distribution of food count by Day of the Week")
legend("topright", levels(nonGUIConsumptions$dowFactored) , fi
ll=colfill)
```

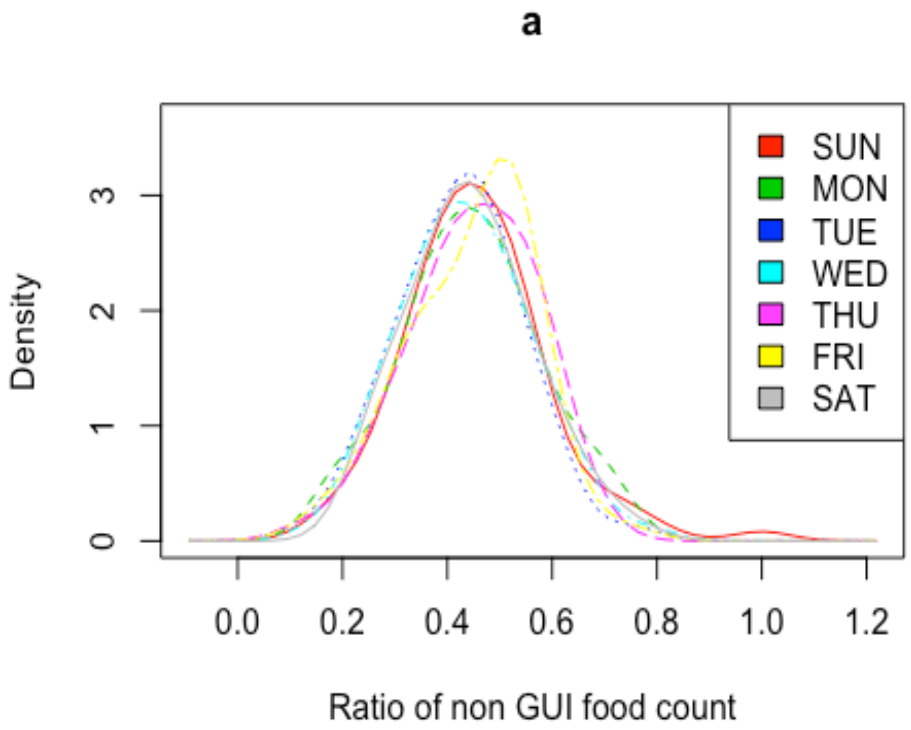



```
sm.density.compare(nonGUIConsumptions$day_fwt, nonGUIConsumpti
ons$dowFactored, xlab="food weight" )
title("Distribution of food weight by Day of the Week")
```

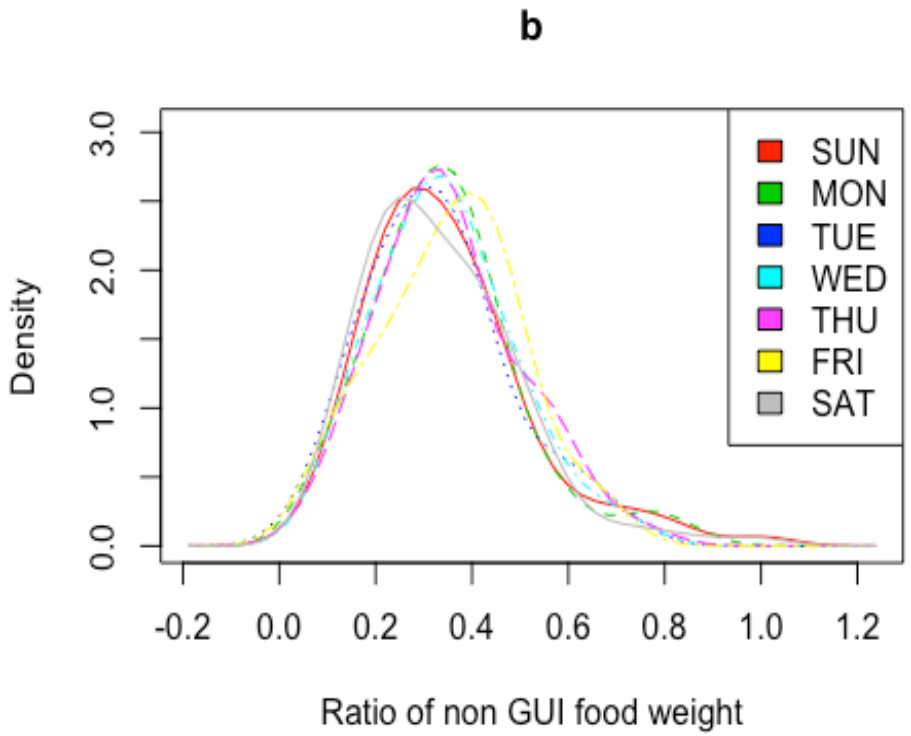
```
legend("topright", levels(nonGUIConsumptions$dowFactored) , fi
ll=colfill)
```




```
sm.density.compare(nonGUIConsumptions$ratio_ct, nonGUIConsumpt
ions$dowFactored, xlab="Ratio of non GUI food count" )
title("Distribution of ratio of non GUI food count by Day of t
he Week")
legend("topright", levels(nonGUIConsumptions$dowFactored) , fi
ll=colfill)
```

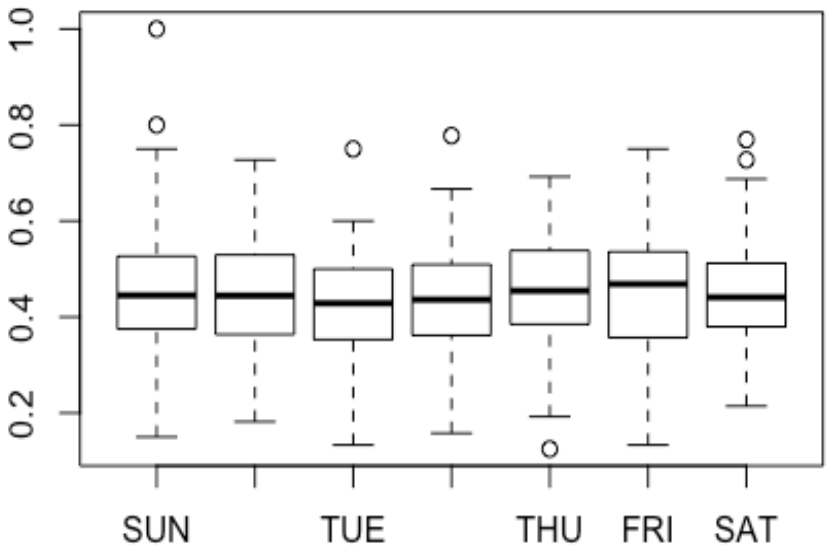



```
sm.density.compare(nonGUIConsumptions$ratio_fwt, nonGUIConsump
tions$dowFactored, xlab="Ratio of non GUI food weight")
title("Distribution of ratio of non GUI food weight by Day of
the Week")
legend("topright", levels(nonGUIConsumptions$dowFactored) , fi
ll=colfill)
```



- **Test for ratio of count**

```
plot(nonGUIConsumptions$DOW, nonGUIConsumptions$ratio_ct)
```




```
print("==== RATIO CT === ")
```

```
## [1] "==== RATIO CT === "
```

```
ratioCtCoeff <- sapply(DOW, function(x, nonGUIConsumptions){
  xInclude <- nonGUIConsumptions[nonGUIConsumptions$DOW == x,
'ratio_ct']
  xExclude <- nonGUIConsumptions[nonGUIConsumptions$DOW != x,
'ratio_ct' ]

  xInclude<- sample(xInclude, 5000, replace = TRUE)
  xExclude<- sample(xExclude, 5000, replace = TRUE)
  print(sprintf("Permutation test = %s", x))
  print(wilcox.test(xInclude, xExclude))
}, nonGUIConsumptions)
```

```
## [1] "Permutation test = SUN"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 12866000, p-value = 0.01113
```

```
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = MON"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 12470000, p-value = 0.8348
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = TUE"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 10924000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = WED"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 11747000, p-value = 1.773e-07
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = THU"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 13373000, p-value = 1.407e-09
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = FRI"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 12913000, p-value = 0.00421
## alternative hypothesis: true location shift is not equal to
0
##
```

```
## [1] "Permutation test = SAT"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 12154000, p-value = 0.01647
## alternative hypothesis: true location shift is not equal to
0

print("=== DAY CT ==== ")

## [1] "=== DAY CT ==== "

ratioCtCoeff <- sapply(DOW, function(x, nonGUIConsumptions){
  xInclude <- nonGUIConsumptions[nonGUIConsumptions$DOW == x,
'day_ct']
  xExclude <- nonGUIConsumptions[nonGUIConsumptions$DOW != x,
'day_ct' ]

  xInclude<- sample(xInclude, 5000, replace = TRUE)
  xExclude<- sample(xExclude, 5000, replace = TRUE)
  print(sprintf("Permutation test = %s", x))
  print(wilcox.test(xInclude, xExclude))
}, nonGUIConsumptions)

## [1] "Permutation test = SUN"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 11038000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = MON"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 12590000, p-value = 0.5321
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = TUE"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 13272000, p-value = 7.967e-08
```

```
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = WED"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 14072000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = THU"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 14092000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = FRI"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 11774000, p-value = 4.618e-07
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = SAT"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 11614000, p-value = 7.766e-10
## alternative hypothesis: true location shift is not equal to
0
```

Test for ratio of food weight

```
print("==== RATIO FWT=====")

## [1] "==== RATIO FWT====="

ratioCtCoeff <- sapply(DOW, function(x, nonGUIConsumptions){
  xInclude <- nonGUIConsumptions[nonGUIConsumptions$DOW == x,
```

```
'ratio_fwt']
  xExclude <- nonGUIConsumptions[nonGUIConsumptions$DOW != x,
'ratio_fwt' ]

  xInclude<- sample(xInclude, 5000, replace = TRUE)
  xExclude<- sample(xExclude, 5000, replace = TRUE)

  print(sprintf("Permutation test = %s", x))
  print(wilcox.test(xInclude, xExclude))
}, nonGUIConsumptions)
```

```
## [1] "Permutation test = SUN"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 12506000, p-value = 0.9688
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = MON"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 12335000, p-value = 0.2539
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = TUE"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 11531000, p-value = 1.93e-11
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = WED"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 12841000, p-value = 0.01831
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = THU"
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 13089000, p-value = 4.479e-05
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = FRI"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 14019000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = SAT"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 12023000, p-value = 0.0009522
## alternative hypothesis: true location shift is not equal to
0

print("=== DAY FWT ==== ")

## [1] "=== DAY FWT ==== "

ratioCtCoeff <- sapply(DOW, function(x, nonGUIConsumptions){
  xInclude <- nonGUIConsumptions[nonGUIConsumptions$DOW == x,
'day_fwt']
  xExclude <- nonGUIConsumptions[nonGUIConsumptions$DOW != x,
'day_fwt' ]

  xInclude<- sample(xInclude, 5000, replace = TRUE)
  xExclude<- sample(xExclude, 5000, replace = TRUE)

  print(sprintf("Permutation test = %s", x))
  print(wilcox.test(xInclude, xExclude))
}, nonGUIConsumptions)

## [1] "Permutation test = SUN"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 10519000, p-value < 2.2e-16
```

```
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = MON"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 13713000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = TUE"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 15005000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = WED"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 14859000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = THU"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 12737000, p-value = 0.1005
## alternative hypothesis: true location shift is not equal to
0
##
## [1] "Permutation test = FRI"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 11359000, p-value = 2.707e-15
## alternative hypothesis: true location shift is not equal to
0
##
```

```
## [1] "Permutation test = SAT"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  xInclude and xExclude
## W = 11247000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to
0
```

- **Daily intake and Snacks**

```
missedFoodAvgDailyIntakeBySubjects <- sqldf("SELECT SUBJECID,
IUNA_NPNS_77FG,
                                CASE WHEN MTYPE IN (6,7,8,11) TH
EN 'Snack' ELSE 'Main meal' END meal_type,
                                SUM(CASE WHEN GUI_CODE IS NULL T
HEN 1 ELSE 0 END)/4.0 non_gui_fq_daily_avg,
                                SUM(CASE WHEN GUI_CODE IS NULL T
HEN FWT ELSE 0 END)/4.0 non_gui_fwt_daily_avg
                                FROM foodDataGUIMapped
                                GROUP BY SUBJECID, IUNA_NPNS_77F
G,
                                CASE WHEN MTYPE IN (6,7,8,11) TH
EN 'Snack' ELSE 'Main meal' END
              ", verbose = TRUE)

missedFoodDailySummary <- as.data.frame( as.matrix(
  aggregate(non_gui_fwt_daily_avg~IUNA_NPNS_77FG+meal_type, da
ta=missedFoodAvgDailyIntakeBySubjects, function(x){c(mean(x),
sd(x), quantile(x,0.50), IQR(x))})
  ))
colnames(missedFoodDailySummary) <- c('IUNA_NPNS_77FG', 'meal_
type', 'avg_di', 'sd_di', 'p50_di', 'iqr_di')


exportSummary <-
merge(
missedFoodDailySummary[missedFoodDailySummary$meal_type == 'Sn
ack', c(1,3,4,5,6)],
missedFoodDailySummary[missedFoodDailySummary$meal_type != 'Sn
ack', c(1,3,4,5,6)],
by= "IUNA_NPNS_77FG",
suffixes = c("_snack", "_main_meal"),
all=TRUE
)
```

```
kable(exportSummary, caption = "Summary stats using daily inta
ke")
```

stats using daily intake

- **Consumption averages**

```
missedFoodMeanIntakeBySubjects <- sqldf("SELECT SUBJECID, SURV
DAY, IUNA_NPNS_77FG,
                                CASE WHEN MTYPE IN (6,7,8,11) TH
EN 'Snack' ELSE 'Main meal' END meal_type,
                                SUM(CASE WHEN GUI_CODE IS NULL T
HEN 1 ELSE 0 END) non_gui_fq_daily_avg,
                                AVG(CASE WHEN GUI_CODE IS NULL T
HEN FWT ELSE 0 END)  non_gui_fwt_daily_avg
                                FROM foodDataGUIMapped
                                GROUP BY SUBJECID, IUNA_NPNS_77F
G, SURVDAY,
                                CASE WHEN MTYPE IN (6,7,8,11) TH
EN 'Snack' ELSE 'Main meal' END
              ")


missedFoodMeanSummary <- as.data.frame( as.matrix(
  aggregate(non_gui_fwt_daily_avg~IUNA_NPNS_77FG+meal_type, da
ta=missedFoodMeanIntakeBySubjects, function(x){c(mean(x), sd(x
), quantile(x,0.50), IQR(x))})
))
colnames(missedFoodDailySummary) <- c('IUNA_NPNS_77FG', 'meal_
type', 'avg_di', 'sd_di', 'p50_di', 'iqr_di')


exportMeanSummary <-
  merge(
    missedFoodMeanSummary[missedFoodMeanSummary$meal_type == '
Snack', c(1,3,4,5,6)],
    missedFoodMeanSummary[missedFoodMeanSummary$meal_type != '
Snack', c(1,3,4,5,6)],
    by= "IUNA_NPNS_77FG",
    suffixes = c("_snack", "_main_meal"),
    all=TRUE
  )

kable(exportMeanSummary, caption = "Summary stats using consum
ption averages")
```

3. Cariogenic Summary

Introduction

This document inlcuded all summary statistics for cariogenic foods.

Loading the mapped data

The same mapped data was used as for the non-covered analysis. The cariogenic food was remaped back to a GUI code.

Libraries and input data

Libraries used and the functions defined for the analysis of the data

```
library(knitr)
library(sqldf)
library(pastecs)

# Read SPSS data
library(foreign)

library(reshape)

# Fancier plots
library(ggplot2)
# Bean plot library
library(beanplot)

if(Sys.info()['sysname'] == "Linux" || Sys.info()['sysname'] =
= "Darwin"){
  setwd("~/Dropbox/iuna-gui/")
  foodDataGUIMapped <- read.csv("foodDataGUIMappedV2.csv")
}else{
  setwd("C:/Users/Michael Crowe/Dropbox/iuna-gui")
  foodDataGUIMapped <- read.csv("foodDataGUIMappedV2.csv")
}

foodDataSPSS <- read.spss(file="npns-food-file-4R.sav")
iunaFoodCodes<- cbind.data.frame(code=seq(1,77), description=l
evels(foodDataSPSS$IUNA_NPNS_77FG))

mappingTable <- sqldf("select IUNA_NPNS_77FG, CMETH, Food_desc
ription_first_first, GUI_CODE, count(*) ct FROM foodDataGUIMap
ped
      GROUP BY IUNA_NPNS_77FG, CMETH, Food_description_first_f
irst, GUI_CODE")
```

- **NPNS F77 Cariogenic food codes**

A list of cariogenic foods was manually identified in the NPNS data.

```
cariogenicIUNAFoodCode <- c(6, 8, 9, 16, 17, 18, 35, 39, 57, 5
8, 59, 66, 68)
kable(iunaFoodCodes[cariogenicIUNAFoodCode,], row.names =FALSE
)
```

code	description
6	RTEBC
8	Biscuits including crackers
9	Cakes, pastries & buns
16	Ice creams
17	Desserts
18	Rice puddings & custards
35	Fruit juices
39	Tinned fruit
57	Sugars, syrups, preserves & sweeteners
58	Chocolate confectionery
59	Non-chocolate confectionery
66	Carbonated beverages
68	Squashes, cordials & fruit juice drinks

- **Constants and functions**

```
convertAggregateToDataframe <- function(aggRst, colNames = c()
) {
  df <- as.data.frame(as.matrix(aggRst))
  if (length(colNames) > 0) {
  colnames(df) <- colNames
  }
  df
  }


# Extends aggreated metrics including P25, P75, P95 and P99
  computeFoodMetrics <- function(x) {
  m  <- stat.desc(x)
  q <- quantile(x, c(0.25, 0.75, 0.95, 0.99))
  names(q) <- c("p25", "p75", "p95", "p99")
  append(m,q)
  }
```

```
# What columns to report in the tables
  baseColSelectionPattern <- "IUNA|nbr.val|FWT.mean$|FQ.mean$|
median|std.dev|p95|max"


  mapSnacks <- function(x) {
  isSnack <- x %in% c(6,7,8,11)
  retVal <- 0
  if (isSnack) {
  retVal <- 1
  }
  retVal
  }

  castMealTypeDF <- function(inputDf) {
  snackIdx <-
  grep("IS_SNACK", colnames(inputDf))
  iunaIdx <- grep("IUNA", colnames(inputDf))

  mainMeal <-
  inputDf[inputDf$IS_SNACK == 0,-c(snackIdx)]
  colnames(mainMeal) <-
  paste(colnames(mainMeal), ".MainMeal", sep = "")
  snacks <-
  inputDf[inputDf$IS_SNACK == 1,-c(snackIdx)]
  colnames(snacks) <- paste(colnames(snacks), ".Snacks", sep =
"")

  df <- cbind(snacks, mainMeal)
  iunaIdx <- grep("IUNA", colnames(df))

  iunCol <- colnames(df)[iunaIdx]
  otherCols <- colnames(df)[-iunaIdx]
  otherCols <- otherCols[order(otherCols)]
  df <- df[, c(iunCol, otherCols)]
  df


  }

  buildCariogenicOnlyDf <- function (df, foodCodes, columPatte
rn) {
  colSelection <-
  grep(columPattern ,names(df))
  retDf <-
  df[df$IUNA_NPNS_77FG %in% foodCodes, colSelection]
  retDf
```

```
  }

  prettyColNamesDf <- function(df){
    colnames(df) <- gsub("\\.|_", " ", colnames(df))
    df
  }

  prettyKable <- function(df, digits=2){
  kable(prettyColNamesDf(df), digits = digits, row.names = FAL
SE )
  }
  n.days = 4

  foodDataGUIMapped$IS_SNACK <-
  sapply(foodDataGUIMapped$MTYPE, mapSnacks)


  foodKSTest <- function(df) {
    testVar <- names(df)[which(names(df) != 'IS_SNACK')]
    x <- df[df$IS_SNACK == 0, testVar]
    y <- df[df$IS_SNACK == 1, testVar]
    if (length(x) < 10 || length(y) < 10) {
    warning(paste("One or both splits are too small ", length(
x), length(y)))
    return(cbind(NA, NA))
    }
    ksResult <- ks.test(x, y, alternative = "two.sided")
    cbind(ksResult$statistic, ksResult$p.value)
  }
```

- **Global statistics**

Computed the number of subjects that consumed either never consumed a cariogenic food or consumed it.

```
foodDataGUIMapped$IS_CARIOG <- foodDataGUIMapped$IUNA_NPNS_77F
G %in% cariogenicIUNAFoodCode

subjectEatingCG <-  unique(foodDataGUIMapped[foodDataGUIMapped
$IS_CARIOG == TRUE, 'SUBJECID'])
length(subjectEatingCG)

## [1] 126

length(unique(foodDataGUIMapped$SUBJECID))

## [1] 126

setdiff(unique(foodDataGUIMapped$SUBJECID), subjectEatingCG)
```

```
## integer(0)

subMealSurvCGSummary <- sqldf("select SUBJECID, MEALNO, SURVDA
Y,
      SUM(CASE WHEN IS_CARIOG THEN 1 ELSE 0 END) CG_CT,
      COUNT(IUNA_NPNS_77FG) CT,
      SUM(CASE WHEN IS_CARIOG THEN FWT ELSE 0 END) CG_FWT,
      SUM(FWT) FTW
      from foodDataGUIMapped
       group by SUBJECID, MEALNO, SURVDAY
      ")


dowSummary <- sqldf("select mtype, dow, avg(cg_ct) cg_ct,
      avg(ct) ct,
      avg(cg_fwt) cg_fwt,
      avg(fwt) fwt
      from
      (select SUBJECID, MTYPE, DOW,
      SUM(CASE WHEN IS_CARIOG THEN 1 ELSE 0 END) CG_CT,
      COUNT(IUNA_NPNS_77FG) CT,
      SUM(CASE WHEN IS_CARIOG THEN FWT ELSE 0 END) CG_FWT,
      SUM(FWT) fwt
      from foodDataGUIMapped
       group by SUBJECID, MTYPE, DOW)
       group by mtype, dow
       order by dow, mtype
      ")

mealtypeCodes <- read.table(file="M_Type.txt", sep = "=")
dowLabel <- c('MON', 'TUE', 'WED', 'THU', 'FRI', 'SAT', 'SUN')
colorCols <- c('red', 'gray')
for(i in 1:9){
png(filename = sprintf("~/%s-meal.png", i), width = 1980, heig
ht = 1024, units = "px")
tmpDf <- dowSummary[dowSummary$MTYPE == i, ]
tmpDf <- as.matrix(sapply(tmpDf$DOW, function(i){tmpDf[tmpDf$D
OW == i, c('cg_ct', 'ct', 'cg_fwt', 'fwt')]}))
colnames(tmpDf) <- c('MON', 'TUE', 'WED', 'THU', 'FRI', 'SAT',
'SUN')
par(mfrow=c(1,2), mar=c(5,4,8,3))
barplot(tmpDf[1:2, ], legend.text = c('cariogenic', 'total'),
xlab="DOW", ylab="AVG #{of components}", ylim = c(0, 8),
        main = "Component count", col=colorCols)
barplot(tmpDf[3:4, ], legend.text = c('cariogenic', 'total'),
xlab="DOW", ylab="AVG weight (g)",  ylim = c(0, 500), main="Da
ily average weights",
        col=colorCols)
```

```
title(outer = TRUE, main = sprintf("Meal Type = %s", mealtypeC
odes[i, 2]), line = -1)
dev.off()
}

for(i in 1:7){
png(filename = sprintf("~/%s-dow.png", i), width = 1980, heigh
t = 1024, units = "px")
tmpDf <- dowSummary[dowSummary$DOW == i, ]
meals <- 1:9
tmpDf <- as.matrix(sapply(meals, function(i){tmpDf[tmpDf$MTYPE
== i, c('cg_ct', 'ct', 'cg_fwt', 'fwt')]}))
colnames(tmpDf) <- mealtypeCodes[1:9, 2]
par(mfrow=c(1,2), mar=c(5,4,8,3))
barplot(tmpDf[1:2, ], legend.text = c('cariogenic', 'total'),
xlab="Meal type", ylab="AVG #{of components}", ylim = c(0, 8),
        main = "Component count", col=colorCols, cex.names = 0
.7)
barplot(tmpDf[3:4, ], legend.text = c('cariogenic', 'total'),
xlab="Meal type", ylab="AVG weight (g)",  ylim = c(0, 500), ma
in="Daily average weights",
        col=colorCols, cex.names = 0.7)
title(outer = TRUE, main = sprintf("DOW = %s", dowLabel[i]), l
ine = -1)
dev.off()
}
```

All subjects during the 4 days consumed at least one of the cariogenic food.

If we look at the number of meals independently from the subject:

```
totalMeals <- nrow(subMealSurvCGSummary)
mealsContainigCG <- nrow(subMealSurvCGSummary[subMealSurvCGSum
mary$CG_CT >0 ,])

prettyKable(cbind( total_meals=totalMeals, meals_containing_ca
riogenic= mealsContainigCG, "ratio_%"=mealsContainigCG/totalMe
als))
```

total meals	meals containing cariogenic	ratio %
2676	1500	0.56

for each meal we look how many food *components / ingredients* were cariogenic.

Looked at how many *components* in a meal.

```
mSummary <- computeFoodMetrics(subMealSurvCGSummary$CT)
prettyKable(cbind(stats=names(mSummary), val=convertAggregateT
oDataframe(mSummary)))
```

stats	V1
nbr.val	2676.00
nbr.null	0.00
nbr.na	0.00
min	1.00
max	14.00
range	13.00
sum	9211.00
median	3.00
mean	3.44
SE.mean	0.04
CI.mean.0.95	0.07
var	3.85
std.dev	1.96
coef.var	0.57
p25	2.00
p75	5.00
p95	7.00
p99	9.00

```
prettyKable(as.data.frame(round(table(subMealSurvCGSummary$CT)
/nrow(subMealSurvCGSummary)*100,2)))
```

Var1	Freq
1	17.08
2	19.73
3	19.92
4	17.00
5	11.70
6	7.06
7	3.96
8	1.98
9	0.82
10	0.41
11	0.19
12	0.07
13	0.04
14	0.04

and how many cariogenic components are in a meal

```
cgSummary <- computeFoodMetrics(subMealSurvCGSummary$CG_CT)
prettyKable(cbind(stats=names(cgSummary), val=convertAggregate
ToDataframe(cgSummary)))
```

stats	V1
nbr.val	2676.00
nbr.null	1176.00
nbr.na	0.00
min	0.00
max	8.00
range	8.00
sum	2169.00
median	1.00
mean	0.81
SE.mean	0.02
CI.mean.0.95	0.03
var	0.84
std.dev	0.92
coef.var	1.13
p25	0.00
p75	1.00
p95	2.00
p99	4.00

```
prettyKable(as.data.frame(round(table(subMealSurvCGSummary$CG_
CT)/nrow(subMealSurvCGSummary)*100,2)))
```

Var1	Freq
0	43.95
1	37.11
2	14.80
3	3.10
4	0.64
5	0.15
6	0.15
7	0.04
8	0.07

```
## As pecerntage
computeFoodMetrics(subMealSurvCGSummary$CG_CT/subMealSurvCGSum
mary$CT*100)

##      nbr.val     nbr.null       nbr.na          min
max
## 2.676000e+03 1.176000e+03 0.000000e+00 0.000000e+00 1.00000
```

```
0e+02
##        range          sum       median         mean      SE
.mean
## 1.000000e+02 6.840909e+04 2.000000e+01 2.556394e+01 5.76054
5e-01
## CI.mean.0.95          var      std.dev     coef.var
p25
## 1.129557e+00 8.880006e+02 2.979934e+01 1.165679e+00 0.00000
0e+00
##          p75          p95          p99
## 5.000000e+01 1.000000e+02 1.000000e+02
```

```
averageFQCariogenicFoodDay <- sqldf("select avg(avg_fq) avg_fq
_consumption_average,
avg(avg_daily) avg_fq_daily_intakes
from
      (select SUBJECID, avg(CG_CT) avg_fq,
      SUM(CG_CT) /4 avg_daily  from (select SUBJECID, SURVDAY,
      SUM(CASE WHEN IS_CARIOG THEN 1 ELSE 0 END) CG_CT,
      COUNT(IUNA_NPNS_77FG) CT,
      SUM(CASE WHEN IS_CARIOG THEN FWT ELSE 0 END) CG_FWT,
      SUM(FWT) FTW
      from foodDataGUIMapped
       group by SUBJECID, SURVDAY ) group by SUBJECID)
      ")

kable(averageFQCariogenicFoodDay, caption = "Summary cariogeni
c frequency - all subjects are
      consumers")
```

Summary cariogenic frequency - all subjects are consumers

avg_fq_consumption_average	avg_fq_daily_intakes
4.303571	3.944444

If we want to look at the subject level

```
prettyKable(sqldf("select CG_CT 'nbr cariogenic components', C
OUNT(DISTINCT SUBJECID) 'nbr subjects', COUNT(*) 'frequency' f
rom subMealSurvCGSummary group by CG_CT"))
```

nbr cariogenic components	nbr subjects	frequency
0	125	1176
1	126	993
2	116	396
3	56	83
4	14	17
5	4	4
6	4	4
7	1	1
8	2	2

```
prettyKable(sqldf("select CG_CT 'nbr cariogenic components', C
T 'nbr components',
                  COUNT(DISTINCT SUBJECID) 'nbr subjects', COU
NT(*) 'freqency'
                  from subMealSurvCGSummary group by CG_CT,CT
                  order by 3 desc,4 desc limit 20"))
```

nbr cariogenic components	nbr components	nbr subjects	freqency
0	2	100	247
0	1	95	325
1	2	93	232
1	4	91	183
0	3	89	232
1	3	88	194
0	4	85	166
1	5	74	117
1	1	73	132
0	5	67	100
2	3	57	94
2	4	55	83
1	6	53	71
2	5	51	84
0	6	43	58
2	2	40	49
2	6	31	39
1	7	29	39
0	7	24	30
3	4	19	21

```
prettyKable(sqldf("select AVG(CG_FWT/FTW)*100 'avg % food weig
ht' from subMealSurvCGSummary "))
```

avg % food weight
7.4

- **Snacks stats**

```
subjectEatingSnacks <-  unique(foodDataGUIMapped[foodDataGUIMa
pped$IS_SNACK == TRUE, 'SUBJECID'])
length(subjectEatingSnacks)

length(unique(foodDataGUIMapped$SUBJECID))

setdiff(unique(foodDataGUIMapped$SUBJECID), subjectEatingSnack
s)

subSnackSummary <- sqldf("select SUBJECID, SURVDAY,
      SUM(CASE WHEN IS_SNACK THEN 1 ELSE 0 END) SK_CT,
      COUNT(SUBJECID) CT,
      SUM(CASE WHEN IS_SNACK THEN FWT ELSE 0 END) CG_FWT,
      SUM(FWT) FTW
      from (
       select SUBJECID, SURVDAY, IS_SNACK, MEALNO, SUM(FWT) FW
T from
       foodDataGUIMapped
       group by SUBJECID, SURVDAY, IS_SNACK, MEALNO
       )
       group by SUBJECID, SURVDAY
      ")

prettyKable(sqldf("select avg(SK_CT) avg_snack_count, avg(SK_C
T*1.0/CT) avg_snack_pc from subSnackSummary"))
```

avg snack count	avg snack pc
2.07	0.36

```
prettyKable(sqldf("select SK_CT AVG_SNACK_COUNT, COUNT(DISTINC
T SUBJECID) 'subject count', COUNT(DISTINCT SUBJECID) / 126.0*
100 'subject %'
      from (select SUBJECID, ROUND(AVG(SK_CT),0) SK_CT from su
bSnackSummary group by SUBJECID)
      group by SK_CT order by SK_CT"))
```

AVG SNACK COUNT	subject count	subject %
0	2	1.59
1	28	22.22

2	54	42.86
3	30	23.81
4	9	7.14
5	3	2.38

- **Aggregates**

Focusing on **consumers only**

Daily Intakes

Standard way to compute the summary, that is summing the intake for each subject over the 4 days and then divide by 4. The summary statistics are computed across the subjects

In the summary we report

- FTW (food weight)
- FQ (frequency) in this specific case the frequency is computed by summing the number of times the appears in the diary and the dividing by four.

IUNANPNS 77FG	FWT nbr val	FWT max	FWT median	FWT mean	FWT std dev	FWT p95	FQ nbr val	FQ max	FQ median	FQ mean	FQ std dev	FQ p95
6	116	73.50	22.62	23.80	14.30	55.25	116	3.50	1.00	0.93	0.49	1.75
8	95	58.50	11.75	15.06	12.40	40.07	95	7.00	0.75	0.89	0.84	2.25
9	55	50.75	9.75	13.47	10.59	30.47	55	1.25	0.25	0.40	0.23	0.82
16	59	59.75	18.00	21.51	14.47	49.90	59	1.25	0.25	0.41	0.25	1.00
17	23	88.50	17.50	24.71	21.91	68.02	23	1.00	0.25	0.39	0.21	0.75
18	15	158.00	23.25	40.88	42.56	118.27	15	1.25	0.25	0.38	0.28	0.90
35	78	380.00	85.00	105.31	81.49	284.37	78	2.25	0.75	0.78	0.50	1.79
39	5	37.50	7.25	14.30	14.33	33.75	5	0.50	0.25	0.30	0.11	0.45
57	71	28.00	4.25	6.48	6.56	23.50	71	2.75	0.75	0.73	0.54	1.75

58	75	40.00	8.25	10.67	8.40	26.80	75	1.50	0.50	0.50	0.29	1.00
59	57	37.50	8.00	10.26	7.27	23.45	57	1.75	0.50	0.49	0.34	1.10
66	20	222.25	50.00	68.84	51.66	187.34	20	2.00	0.25	0.42	0.41	0.81
68	84	550.00	60.88	83.61	93.08	226.19	84	5.00	1.00	1.21	0.91	2.75

```
bySubject <-
  aggregate(FWT ~ SUBJECID + IUNA_NPNS_77FG, data = foodDataGU
IMapped, function(x) {
  c(sum(x) / n.days, length(x) / n.days)
  })
  bySubject <-
  convertAggregateToDataframe(bySubject, c("SUBJECID", "IUNA_N
PNS_77FG", "FWT", "FQ"))

  dailyIntakeSummary <-
  convertAggregateToDataframe(aggregate(cbind(FWT, FQ) ~ IUNA_
NPNS_77FG, data =
  bySubject, computeFoodMetrics))
  dailyIntakeSummaryCariogenicOnly <- buildCariogenicOnlyDf(da
ilyIntakeSummary, cariogenicIUNAFoodCode, baseColSelectionPatt
ern)
  prettyKable(dailyIntakeSummaryCariogenicOnly, digits = 2)
```

- **Meal level aggregates**

Here we repeat the same stats but we also split between Snacks and Main meals

```
  bySubject <-
    aggregate(FWT ~ SUBJECID + IUNA_NPNS_77FG + IS_SNACK, data
= foodDataGUIMapped, function(x) {
    c(sum(x) / n.days, length(x) / n.days)
    })
    bySubject <-
    convertAggregateToDataframe(bySubject, c("SUBJECID", "IUNA
_NPNS_77FG",
    "IS_SNACK","FWT", "FQ"))

    dailyIntakeMealTypeFWTSummary <-
    convertAggregateToDataframe(aggregate(FWT ~ IUNA_NPNS_77FG
+ IS_SNACK, data = bySubject, computeFoodMetrics))

## Warning in qt((0.5 + p/2), (Nbrval - 1)): NaNs produced

## Warning in qt((0.5 + p/2), (Nbrval - 1)): NaNs produced

## Warning in qt((0.5 + p/2), (Nbrval - 1)): NaNs produced
```

```
## Warning in qt((0.5 + p/2), (Nbrval - 1)): NaNs produced

## Warning in qt((0.5 + p/2), (Nbrval - 1)): NaNs produced

## Warning in qt((0.5 + p/2), (Nbrval - 1)): NaNs produced

## Warning in qt((0.5 + p/2), (Nbrval - 1)): NaNs produced

## Warning in qt((0.5 + p/2), (Nbrval - 1)): NaNs produced
```

```
    dailyIntakeMealTypeFWTSummaryCariogenicOnly <-
    buildCariogenicOnlyDf(
    dailyIntakeMealTypeFWTSummary, cariogenicIUNAFoodCode,
    paste("IS_SNACK|", baseColSelectionPattern, sep = "")
    )


    dailyIntakeMealTypeFWTSummaryCariogenicOnlyRotated <-
    castMealTypeDF(dailyIntakeMealTypeFWTSummaryCariogenicOnly
)

    prettyKable(dailyIntakeMealTypeFWTSummaryCariogenicOnlyRot
ated, digits = 2)
```

IUNA NPNS 77FG Snacks	IUNA NPNS 77FG MainMeal	FWT mean MainMeal	FWT mean Snacks	FWT nbr val MainMeal	FWT nbr val Snacks	FWT std dev MainMeal	FWT std dev Snacks
6	6	22.88	8.04	114	19	13.44	4.71
8	8	6.55	12.69	48	88	5.17	10.44
9	9	11.23	10.28	33	36	9.00	8.57
16	16	18.48	17.32	34	37	12.46	9.79
17	17	21.72	18.40	16	12	21.94	18.87
18	18	32.18	37.04	11	7	32.89	24.03
35	35	89.00	64.00	70	31	65.00	51.11
39	39	17.12	3.00	4	1	14.86	NA
57	57	5.91	3.61	62	26	5.97	2.73
58	58	8.16	9.35	27	62	5.11	7.45
59	59	7.85	9.03	25	43	5.05	6.95
66	66	65.67	48.97	15	8	45.26	22.76
68	68	61.11	40.33	74	62	68.86	45.41

```
    dailyIntakeMealTypeFQSummary <-
    convertAggregateToDataframe(aggregate(FQ ~ IUNA_NPNS_77FG
+ IS_SNACK, data = bySubject, computeFoodMetrics))

## Warning in qt((0.5 + p/2), (Nbrval - 1)): NaNs produced
```

```
## Warning in qt((0.5 + p/2), (Nbrval - 1)): NaNs produced

## Warning in qt((0.5 + p/2), (Nbrval - 1)): NaNs produced

## Warning in qt((0.5 + p/2), (Nbrval - 1)): NaNs produced

## Warning in qt((0.5 + p/2), (Nbrval - 1)): NaNs produced

## Warning in qt((0.5 + p/2), (Nbrval - 1)): NaNs produced

## Warning in qt((0.5 + p/2), (Nbrval - 1)): NaNs produced

## Warning in qt((0.5 + p/2), (Nbrval - 1)): NaNs produced
```

```
    dailyIntakeMealTypeFQSummaryCariogenicOnly <-
    buildCariogenicOnlyDf(
    dailyIntakeMealTypeFQSummary,
    cariogenicIUNAFoodCode,
    paste("IS_SNACK|", baseColSelectionPattern, sep = "")
    )


    dailyIntakeMealTypeFQSummaryCariogenicOnlyRotated <-
    castMealTypeDF(dailyIntakeMealTypeFQSummaryCariogenicOnly)

    prettyKable(dailyIntakeMealTypeFQSummaryCariogenicOnlyRota
ted, digits = 2)
```

IUNA NPNS 77FG Snacks	IUNA NPNS 77FG MainMeal	FQ max MainMeal	FQ mean MainMeal	FQ mean Snacks	FQ nbr val MainMeal	FQ nbr val Snacks	FQ std dev MainMeal	FQ std dev Snacks
6	6	3.50	0.89	0.30	114	19	0.48	0.10
8	8	3.75	0.47	0.70	48	88	0.54	0.52
9	9	1.00	0.33	0.31	33	36	0.17	0.16
16	16	1.00	0.36	0.32	34	37	0.21	0.14
17	17	1.00	0.36	0.27	16	12	0.22	0.07
18	18	1.00	0.32	0.32	11	7	0.23	0.12
35	35	1.75	0.69	0.42	70	31	0.38	0.24
39	39	0.50	0.31	0.25	4	1	0.12	NA
57	57	2.25	0.69	0.36	62	26	0.44	0.16
58	58	0.75	0.37	0.44	27	62	0.14	0.25
59	59	1.00	0.38	0.42	25	43	0.19	0.28
66	66	1.25	0.40	0.31	15	8	0.28	0.18
68	68	3.00	0.96	0.49	74	62	0.73	0.34

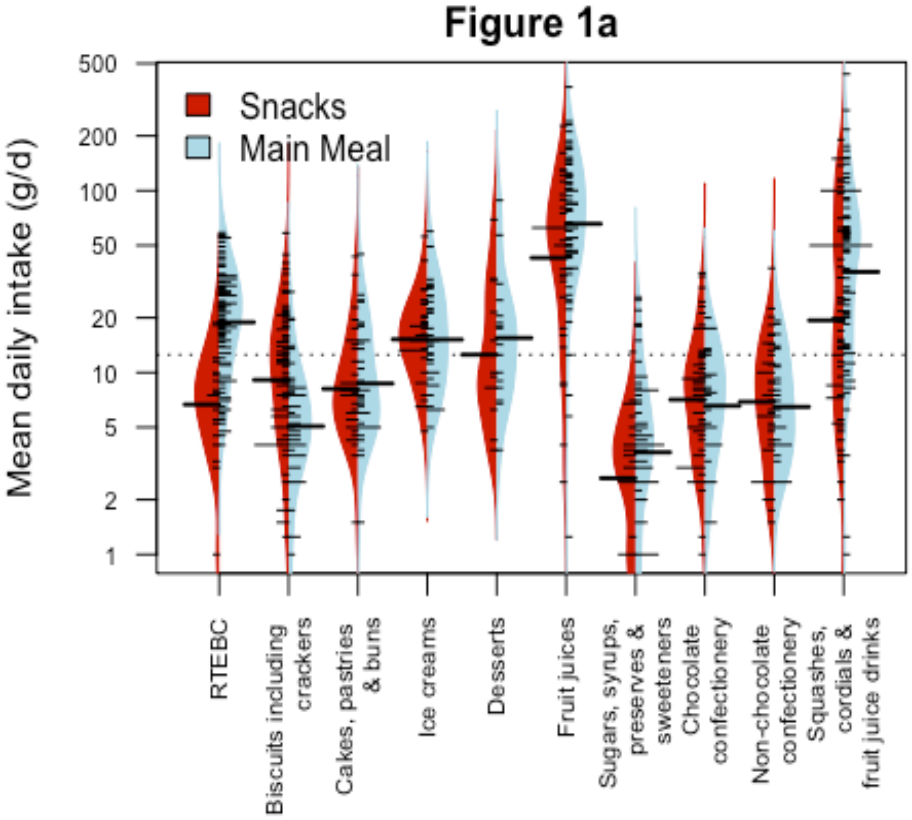
Bean Plots

```
bySubjectSnackReporting <- merge(bySubject, iunaFoodCodes, all
.x = TRUE, by.x="IUNA_NPNS_77FG", by.y = 'code' )
bySubjectSnackReporting$MEAL_TYPE <- sapply(bySubjectSnackRepo
rting$IS_SNACK,

function(x){ retVal<- 'Snack'if(x == 0){retVal <- 'Main Meal'}
retVal})

codesToExculde <- c(18, 39, 66)
kable(iunaFoodCodes[codesToExculde,], row.names = FALSE)
```

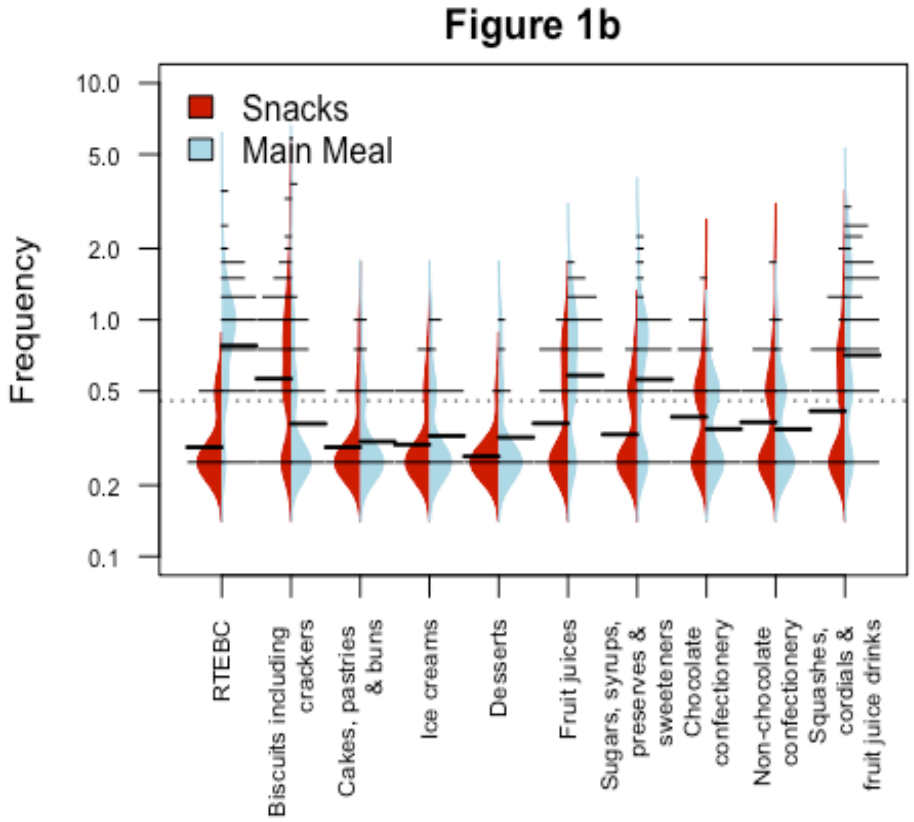
code	description
18	Rice puddings & custards
39	Tinned fruit
66	Carbonated beverages

```
codesForPlot <- cariogenicIUNAFoodCode[(! cariogenicIUNAFoodCo
de %in% codesToExculde) ]

fwt <- bySubjectSnackReporting[bySubjectSnackReporting$IUNA_NP
NS_77FG %in% codesForPlot,
                               c('FWT', 'IUNA_NPNS_77FG', 'IS_
SNACK')]

fq <- bySubjectSnackReporting[bySubjectSnackReporting$IUNA_NPN
S_77FG %in% codesForPlot,
                              c('FQ', 'IUNA_NPNS_77FG', 'IS_SN
ACK')]


beanPlotXLabel <-  c(
  "RTEBC",
  "Biscuits including
   crackers",
  "Cakes, pastries
  & buns",
  "Ice creams",
  "Desserts",
  "Fruit juices",
  "Sugars, syrups,
  preserves &
  sweeteners",
```

```
  "Chocolate
  confectionery",
  "Non-chocolate
  confectionery",
  "Squashes,
   cordials &
   fruit juice drinks"
)
defaultMar <- par()$mar
par(mar=c(6,4,2,2), mfrow=c(1,1))
beanplot(FWT~reorder(IS_SNACK, FWT, mean)*IUNA_NPNS_77FG, data
=fwt,
         side = "b",
         col = list("red3", "lightblue"),
         border = c( "red3", "lightblue"),
         ylim = c(1, 400),
         names = beanPlotXLabel, #iunaFoodCodes$description[co
desForPlot],
         main="Figure 1a",
         ylab="Mean daily intake (g/d)",
         xlab="",
         las=2,
         cex.axis=0.7,
         bw="nrd",
         what = c(1,1,1,1),
         log = "y"

         )
legend("topleft", bty="n", c('Snacks', 'Main Meal'), fill= c("
red3", "lightblue"))
```




```
beanplot(FQ~reorder(IS_SNACK, FQ, mean)*IUNA_NPNS_77FG, data=f
q,
          side = "b",
          col = list("red3", "lightblue"),
          border = c( "red3", "lightblue"),
          ylim = c(0.1, 10),
          names = beanPlotXLabel, #substr(iunaFoodCodes$descrip
tion[codesForPlot], 1, 10),
          main="Figure 1b",
          ylab="Frequency",
          xlab="",
          las=2,
          cex.axis=0.7,
          bw="nrd0",
          what=c(1,1,1,1),
          log = "y"

          )
legend("topleft", bty="n", c('Snacks', 'Main Meal'), fill= c("
red3", "lightblue"))
```




```
par(defaultMar)
```

```
## NULL
```

```
par(mfrow=c(1,1))
```

- **Consumption averages**

Aggregated in a different way: first, aggregate across each subject and each survey day and compute average FTW and frequency, then average across each subject and all 4 survey days and finally look at the stats from these aggregates.

- FWT in here represent the closed food weight at which that food was consumed.
- FQ is the average frequency per day meaning how many times on average the food is consumed on a given day.

```
bySubject <-

 aggregate(FWT ~ SUBJECID + IUNA_NPNS_77FG + SURVDAY, data = f
oodDataGUIMapped, function(x) {
  c(mean(x), length(x))
  })

bySubject <- aggregate(FWT ~ SUBJECID + IUNA_NPNS_77FG , data
= bySubject, mean)

  bySubject <-
  convertAggregateToDataframe(bySubject, c("SUBJECID", "IUNA_N
PNS_77FG", "FWT", "FQ"))

  avgConsSummary <-
  convertAggregateToDataframe(aggregate(cbind(FWT, FQ) ~ IUNA_
NPNS_77FG, data =
  bySubject, computeFoodMetrics))
  avgConsSummaryCariogenicOnly <-
  buildCariogenicOnlyDf(avgConsSummary,
                        cariogenicIUNAFoodCode,
                        baseColSelectionPattern)
  prettyKable(avgConsSummaryCariogenicOnly, digits = 2)
```

IUNA NPNS 77FG	FWT nbr val	FWT max	FWT median	FWT mean	FWT std dev	FQ nbr val	FQ max	FQ median	FQ mean	FQ std dev
6	116	58.75	24.94	26.29	9.99	116	3.50	1.0	1.21	0.38
8	95	56.50	16.50	17.87	9.36	95	7.00	1.0	1.45	0.77
9	55	107.00	30.00	33.38	19.00	55	3.00	1.0	1.11	0.36
16	59	120.00	53.00	54.10	24.36	59	2.00	1.0	1.06	0.20
17	23	277.00	57.00	65.27	55.45	23	2.00	1.0	1.11	0.30
18	15	190.00	93.00	94.51	55.05	15	1.67	1.0	1.04	0.17
35	78	278.00	131.50	141.23	67.70	78	3.00	1.0	1.25	0.44
39	5	150.00	29.00	55.20	59.11	5	1.00	1.0	1.00	0.00
57	71	50.00	8.00	9.82	8.56	71	2.75	1.0	1.24	0.41
58	75	42.00	19.50	20.73	9.01	75	5.00	1.0	1.28	0.66

59	57	90.00	17.50	24.17	18.03	57	2.33	1.0	1.20	0.38
66	20	250.00	181.00	174.21	50.15	20	3.00	1.0	1.20	0.52
68	84	284.00	47.38	80.12	72.77	84	5.00	1.5	1.66	0.75

- **Meal level aggregates**

Repeated the same statistics but also split between Snacks and Main meals

```
  bySubject <-
    aggregate(FWT ~ SUBJECID + IUNA_NPNS_77FG + IS_SNACK + SUR
VDAY, data = foodDataGUIMapped, function(x) {
    c(mean(x), length(x))
    })

   bySubject <- aggregate(FWT ~ SUBJECID + IUNA_NPNS_77FG + IS
_SNACK , data = bySubject, mean)
    bySubject <-
    convertAggregateToDataframe(bySubject, c("SUBJECID", "IUNA
_NPNS_77FG",
    "IS_SNACK","FWT", "FQ"))

bySubjectSnackReporting <- merge(bySubject, iunaFoodCodes, all
.x = TRUE, by.x="IUNA_NPNS_77FG", by.y = 'code' )
bySubjectSnackReporting$MEAL_TYPE <- sapply(bySubjectSnackRepo
rting$IS_SNACK,
                                            function(x){
                                              retVal<- 'Snack'
                                              if(x == 0){retVa
l <- 'Main Meal'}
                                              retVal})
```

- **KS Tests**

```
bySubjectCarioOnly <-
bySubject[bySubject$IUNA_NPNS_77FG %in% cariogenicIUNAFoodCode
,]

#by(bySubjectCarioOnly[c('FQ', 'IS_SNACK')], bySubjectCarioOnl
y[, 'IUNA_NPNS_77FG'], foodKSTest)
#by(bySubjectCarioOnly[, c('FWT', 'IS_SNACK')], bySubjectCario
Only[, 'IUNA_NPNS_77FG'], foodKSTest)

ksTestResults <-
data.frame(matrix(
sapply(cariogenicIUNAFoodCode, function(fcode) {
fq <-
foodKSTest(bySubjectCarioOnly[bySubjectCarioOnly$IUNA_NPNS_77F
```

```
G == fcode, c('FQ', 'IS_SNACK')])
fwt <-
foodKSTest(bySubjectCarioOnly[bySubjectCarioOnly$IUNA_NPNS_77F
G == fcode, c('FWT', 'IS_SNACK')])
cbind(fcode, fq, fwt)
}),
ncol = 5,
byrow = TRUE
))

colnames(ksTestResults) <-
c('IUNA_NPNS_77FG', 'D.FQ', 'p.value.FQ', 'D.FWT', 'p.value.FW
T')

prettyKable(ksTestResults, digits = 4)
```

IUNA NPNS 77FG	D FQ	p value FQ	D FWT	p value FWT
6	0.2368	0.3204	0.1579	0.8116
8	0.2008	0.1635	0.1477	0.5068
9	0.0354	1.0000	0.1414	0.8813
16	0.0318	1.0000	0.2568	0.1932
17	0.0417	1.0000	0.1875	0.9694
18	NA	NA	NA	NA
35	0.1461	0.7489	0.2009	0.3510
39	NA	NA	NA	NA
57	0.2295	0.2894	0.2481	0.2093
58	0.1147	0.9656	0.0938	0.9964
59	0.1526	0.8555	0.1321	0.9455
66	NA	NA	NA	NA
68	0.4438	0.0000	0.1125	0.7870

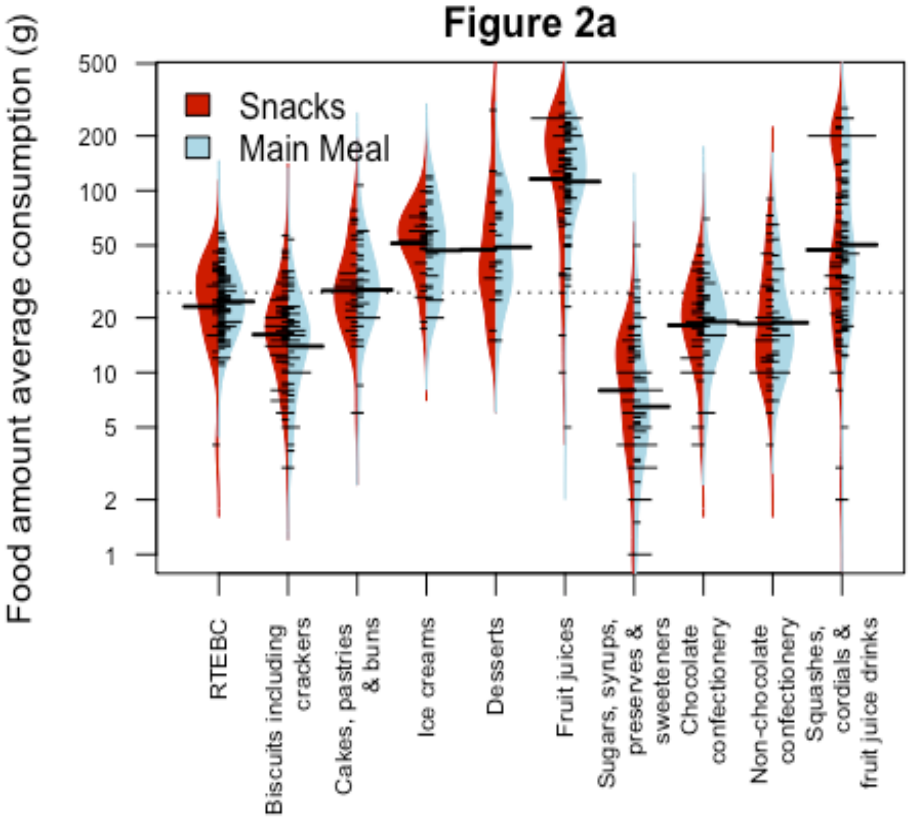
- **Bean plots**

Excluded the following food codes from the bean plots because there were not enough observations:

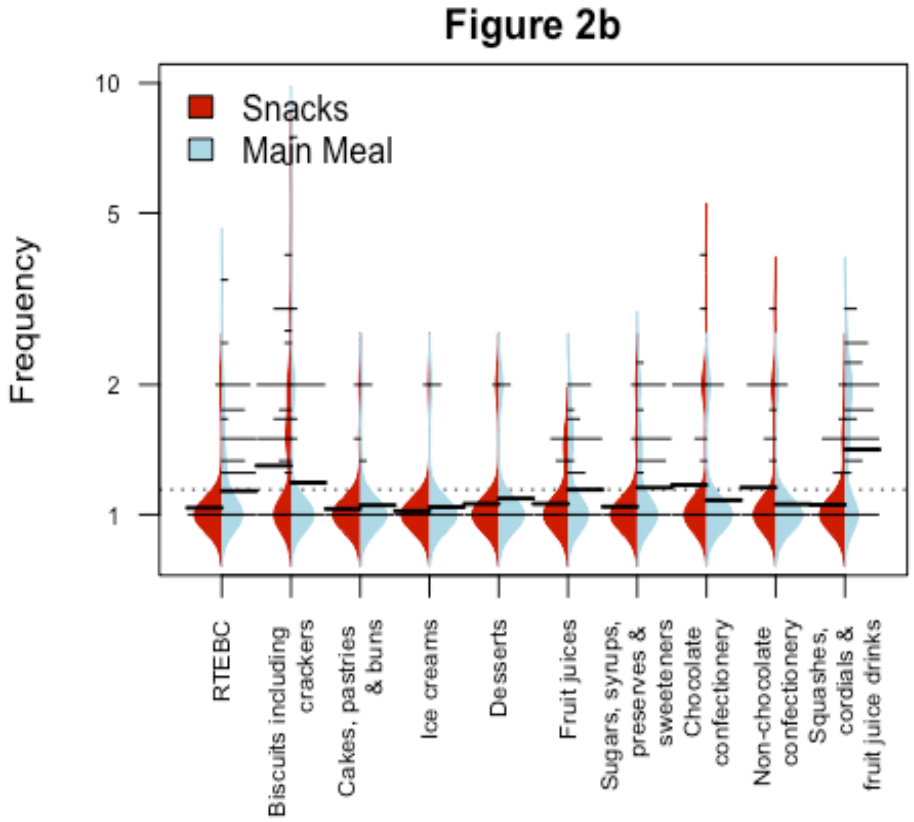
```
codesToExculde <- c(18, 39, 66)
kable(iunaFoodCodes[codesToExculde,], row.names = FALSE)
```

code	description
18	Rice puddings & custards
39	Tinned fruit

```
codesForPlot <- cariogenicIUNAFoodCode[(! cariogenicIUNAFoodCo
de %in% codesToExculde) ]

fwt <- bySubjectSnackReporting[bySubjectSnackReporting$IUNA_NP
NS_77FG %in% codesForPlot,
                               c('FWT', 'IUNA_NPNS_77FG', 'IS_
SNACK')]

fq <- bySubjectSnackReporting[bySubjectSnackReporting$IUNA_NPN
S_77FG %in% codesForPlot,
                              c('FQ', 'IUNA_NPNS_77FG', 'IS_SN
ACK')]

defaultMar <- par()$mar
par(mar=c(6,4,2,2), mfrow=c(1,1))


beanplot(FWT~reorder(IS_SNACK, FWT, mean)*IUNA_NPNS_77FG, data
=fwt,
         side = "b",
         col = list("red3", "lightblue"),
         border = c( "red3", "lightblue"),
         ylim = c(1, 400),
         names = beanPlotXLabel, #substr(iunaFoodCodes$descrip
tion[codesForPlot], 1, 10),
         main="Figure 2a",
         ylab="Food amount average consumption (g)",
         xlab="",
         las=2,
         cex.axis=0.7,
         bw="nrd",
         what = c(1,1,1,1),
         log = "y"
         )


legend("topleft", bty="n", c('Snacks', 'Main Meal'), fill= c("
red3", "lightblue"))
```




```
par(mar=c(6,4,2,2), mfrow=c(1,1))
beanplot(FQ~reorder(IS_SNACK, FQ, mean)*IUNA_NPNS_77FG, data=f
q,
         side = "b",
         col = list("red3", "lightblue"),
         border = c( "red3", "lightblue"),
         ylim = c(0.8, 10),
         names = beanPlotXLabel, #substr(iunaFoodCodes$descrip
tion[codesForPlot], 1, 10),
         main="Figure 2b",
         ylab="Frequency",
         xlab="",
         las=2,
         cex.axis=0.7,
         bw="nrd0",
         what=c(1,1,1,1),
        log = "y"
       )
legend("topleft", bty="n", c('Snacks', 'Main Meal'), fill= c("
red3", "lightblue"))
```




```
    prettyKable(avgConsMealTypeFWTSummaryCariogenicOnlyRotated
, digits = 2)
```

IUNA NPNS 77FG Snacks	IUNA NPNS 77FG MainMeal	FWT max MainMeal	FWT max Snacks	FWT mean MainMeal	FWT mean Snacks	FWT nbr val MainMeal	FWT nbr val Snacks	FWT std dev MainMeal	FWT std dev Snacks
6	6	58.75	46.00	26.39	26.16	114	19	10.36	11.36
8	8	54.00	56.50	16.47	18.53	48	88	9.54	9.79
9	9	107.00	78.00	33.07	32.21	33	36	19.37	17.40
16	16	120.00	99.00	53.88	55.00	34	37	28.90	18.58
17	17	123.00	277.00	57.10	68.12	16	12	31.49	73.28
18	18	222.00	190.00	95.73	114.36	11	7	60.37	69.89
35	35	268.12	302.67	132.85	150.34	70	31	65.17	84.73
39	39	150.00	12.00	66.00	12.00	4	1	62.29	NA
57	57	50.00	27.00	9.26	9.97	62	26	8.79	6.06
58	58	70.00	50.00	21.89	20.91	27	62	12.91	10.29
59	59	65.00	90.00	22.34	24.07	25	43	14.80	19.41
66	66	250.00	210.00	168.34	163.12	15	8	55.56	45.97
68	68	284.00	250.00	76.83	81.40	74	62	70.36	78.56

```
    avgConsMealTypeFQSummary <-
    convertAggregateToDataframe(aggregate(FQ ~ IUNA_NPNS_77FG
+ IS_SNACK, data = bySubject, computeFoodMetrics))

    avgConsMealTypeFQSummaryCariogenicOnly <-
    buildCariogenicOnlyDf(
    avgConsMealTypeFQSummary,
    cariogenicIUNAFoodCode,
    paste("IS_SNACK|", baseColSelectionPattern, sep = "")
    )


    avgConsMealTypeFQSummaryCariogenicOnlyRotated <-
    castMealTypeDF(avgConsMealTypeFQSummaryCariogenicOnly)

    prettyKable(avgConsMealTypeFQSummaryCariogenicOnlyRotated,
digits = 2)
```

IUNA NPNS 77FG	IUNA NPNS 77FG	FQ max MainMeal	FQ max Snacks	FQ mean MainMeal	FQ mean Snacks	FQ nbr val MainMeal	FQ nbr val Snacks	FQ std dev MainMeal	FQ std dev Snacks

Snacks	MainMeal								
6	6	3.50	2.0	1.17	1.05	114	19	0.37	0.23
8	8	7.50	6.5	1.33	1.42	48	88	1.01	0.80
9	9	2.00	2.0	1.07	1.04	33	36	0.25	0.18
16	16	2.00	2.0	1.06	1.03	34	37	0.24	0.16
17	17	2.00	2.0	1.12	1.08	16	12	0.34	0.29
18	18	1.33	1.0	1.03	1.00	11	7	0.10	0.00
35	35	2.00	1.5	1.17	1.07	70	31	0.30	0.17
39	39	1.00	1.0	1.00	1.00	4	1	0.00	NA
57	57	2.25	2.0	1.19	1.06	62	26	0.34	0.22
58	58	2.00	4.0	1.11	1.25	27	62	0.32	0.56
59	59	2.00	3.0	1.08	1.21	25	43	0.28	0.44
66	66	2.00	1.0	1.15	1.00	15	8	0.35	0.00
68	68	3.00	2.0	1.51	1.07	74	62	0.55	0.19

- **GUI Coverage**

```
guiCoverage <-     sqldf("select IUNA_NPNS_77FG,
    --    100.0*SUM(case when (GUI_CODE) is null then 1 else 0
end)/count(*) uncovered,
          100.0*SUM(case when (GUI_CODE) is null then 0 else 1
end)/count(*) covered
          from foodDataGUIMapped
          group by 1
          ")

guiMapping <- sqldf("select distinct IUNA_NPNS_77FG, GUI_CODE
from foodDataGUIMapped order by IUNA_NPNS_77FG")


#prettyKable(guiMapping)
```

Final Tables

Using daily intakes

IUNA NPNS 77FG	GUI CODE	FWT nbr val	FWT mean	FWT std dev	FQ mean	FQ std dev	FWT mean MainMeal	FWT mean Snacks	FWT nbr val MainMeal	FWT nbr val Snacks	FWT std dev MainMeal	FWT std dev Snacks	FQ mean MainMeal	FQ mean Snacks	FQ std dev MainMeal	FQ std dev Snacks
RTEBC	NA	116	24	14	0.9	0.5	23	8	114	19	13	5	0.9	0.3	0.5	0.1
Biscuits including crackers	C25g	95	15	12	0.9	0.8	7	13	48	88	5	10	0.5	0.7	0.5	0.5
Cakes, pastries & buns	C25g	55	13	11	0.4	0.2	11	10	33	36	9	9	0.3	0.3	0.2	0.2
Ice creams	NA	59	22	14	0.4	0.3	18	17	34	37	12	10	0.4	0.3	0.2	0.1
Desserts	C25g	23	25	22	0.4	0.2	22	18	16	12	22	19	0.4	0.3	0.2	0.1
Rice puddings & custards	NA	15	41	43	0.4	0.3	32	37	11	7	33	24	0.3	0.3	0.2	0.1
Fruit juices	NA	78	105	81	0.8	0.5	89	64	70	31	65	51	0.7	0.4	0.4	0.2
Tinned fruit	NA	5	14	14	0.3	0.1	17	3	4	1	15	NA	0.3	0.2	0.1	NA
Sugars, syrups, preserves & sweeteners	NA	71	6	7	0.7	0.5	6	4	62	26	6	3	0.7	0.4	0.4	0.2

Chocolate confectionery	C25g	75	11	8	0.5	0.3	8	9	27	62	5	7	0.4	0.4	0.1	0.2
Non-chocolate confectionery	C25h	57	10	7	0.5	0.3	8	9	25	43	5	7	0.4	0.4	0.2	0.3
Carbonated beverages	C25m	20	69	52	0.4	0.4	66	49	15	8	45	23	0.4	0.3	0.3	0.2
Squashes, cordials & fruit juice drinks	C25l	84	84	93	1.2	0.9	61	40	74	62	69	45	1.0	0.5	0.7	0.3
Squashes, cordials & fruit juice drinks	C25m	84	84	93	1.2	0.9	61	40	74	62	69	45	1.0	0.5	0.7	0.3

Using consumption averages

IUNA NPNS 77FG	GUI CODE	FWT nbr val	FWT mean	FWT std dev	FQ mean	FQ std dev	FWT mean MainMeal	FWT mean Snacks	FWT nbr val MainMeal	FWT nbr val Snacks	FWT std dev MainMeal	FWT std dev Snacks	FQ mean MainMeal	FQ mean Snacks	FQ std dev MainMeal	FQ std dev Snacks
RTEBC	NA	116	26	10	1.2	0.4	26	26	114	19	10	11	1.2	1.1	0.4	0.2
Biscuits including crackers	C25g	95	18	9	1.4	0.8	16	19	48	88	10	10	1.3	1.4	1.0	0.8
Cakes, pastries & buns	C25g	55	33	19	1.1	0.4	33	32	33	36	19	17	1.1	1.0	0.2	0.2
Ice creams	NA	59	54	24	1.1	0.2	54	55	34	37	29	19	1.1	1.0	0.2	0.2
Desserts	C25g	23	65	55	1.1	0.3	57	68	16	12	31	73	1.1	1.1	0.3	0.3
Rice puddings & custards	NA	15	95	55	1.0	0.2	96	114	11	7	60	70	1.0	1.0	0.1	0.0
Fruit juices	NA	78	141	68	1.3	0.4	133	150	70	31	65	85	1.2	1.1	0.3	0.2
Tinned fruit	NA	5	55	59	1.0	0.0	66	12	4	1	62	NA	1.0	1.0	0.0	NA
Sugars, syrups, preserves & sweeteners	NA	71	10	9	1.2	0.4	9	10	62	26	9	6	1.2	1.1	0.3	0.2

Chocolate confectionery	C25g	75	21	9	1.3	0.7	22	21	27	62	13	10	1.1	1.2	0.3	0.6
Non-chocolate confectionery	C25h	57	24	18	1.2	0.4	22	24	25	43	15	19	1.1	1.2	0.3	0.4
Carbonated beverages	C25m	20	174	50	1.2	0.5	168	163	15	8	56	46	1.1	1.0	0.4	0.0
Squashes, cordials & fruit juice drinks	C25l	84	80	73	1.7	0.8	77	81	74	62	70	79	1.5	1.1	0.6	0.2
Squashes, cordials & fruit juice drinks	C25m	84	80	73	1.7	0.8	77	81	74	62	70	79	1.5	1.1	0.6	0.2

- **Consumption averages all cariogenic food excluding RTEBC, Fruit Juice, Tin fruit**

```
cariogenicConsumpations <-
foodDataGUIMapped[foodDataGUIMapped$IUNA_NPNS_77FG %in% setdif
f(cariogenicIUNAFoodCode, c(6, 35,39)),]

  bySubject <-
  aggregate(FWT ~ SUBJECID + SURVDAY, data = cariogenicConsump
ations, function(x) {
  c(mean(x), length(x))
  })

  bySubject <-
  aggregate(FWT ~ SUBJECID  , data = bySubject, mean)

  bySubject <-
  convertAggregateToDataframe(bySubject, c("SUBJECID", "FWT",
"FQ"))

  avgConsSummaryTotal <-
  convertAggregateToDataframe(aggregate(cbind(FWT, FQ) ~ 1 , d
ata =
  bySubject, computeFoodMetrics))

  summaryTable <- t(avgConsSummaryTotal)
  colnames(summaryTable) <- c('value')
  kable(summaryTable)
```

	value
FWT.nbr.val	126.0000000
FWT.nbr.null	0.0000000
FWT.nbr.na	0.0000000
FWT.min	2.5000000
FWT.max	133.4375000
FWT.range	130.9375000
FWT.sum	5102.7240079
FWT.median	35.4791667
FWT.mean	40.4978096
FWT.SE.mean	2.1991915
FWT.CI.mean.0.95	4.3524730
FWT.var	609.3918769
FWT.std.dev	24.6858639

FWT.coef.var	0.6095605
FWT.p25	22.7590278
FWT.p75	52.0750000
FWT.p95	86.7375000
FWT.p99	129.7705357
FQ.nbr.val	126.0000000
FQ.nbr.null	0.0000000
FQ.nbr.na	0.0000000
FQ.min	1.0000000
FQ.max	7.5000000
FQ.range	6.5000000
FQ.sum	397.8333333
FQ.median	3.0000000
FQ.mean	3.1574074
FQ.SE.mean	0.1268921
FQ.CI.mean.0.95	0.2511352
FQ.var	2.0288025
FQ.std.dev	1.4243604
FQ.coef.var	0.4511171
FQ.p25	2.0000000
FQ.p75	4.0000000
FQ.p95	5.4375000
FQ.p99	7.3750000

4. Association Analysis

Introduction

Goal is to explore what foods are consumed together in a meal.

The focus is cariogenic foods.

Some key points

1. Identify the components that goes into a meal
2. Compare meals when the components are describe using the GUI coding vs using the IUNA FG 77 coding
3. Understand the combinations of components that characterize the most common meals.
4. How each cariogenic food interacts with non-cariogenic and other cariogenic food.

Functions and data loading

```
library(sqldf)

library(pastecs)

library(foreign)

source('cariogenic-codes.R')


getSplitWeightGUI <- function(consmuptions, id, day) {
  if (nrow(consmuptions) == 0) {
    print(id)
    print(day)
    stop(consmuptions)
  }
  agg <- aggregate(FWT ~ is.na(GUI_CODE), consmuptions, sum)
  nonguiIdx <- which(agg$`is.na(GUI_CODE)`)
  guiIdx <- which(agg$`is.na(GUI_CODE)` == FALSE)
  nonGuiFWT <- NA
  if (length(nonguiIdx) > 0) {
    nonGuiFWT <- unlist(agg$FWT[nonguiIdx])
  }
  guiFWT <- NA
  if (length(guiIdx) > 0) {
    guiFWT <- unlist(agg$FWT[guiIdx])
  }

  list(gui = guiFWT, nonGui = nonGuiFWT)
```

```
}

  prettyColNamesDf <- function(df){
    colnames(df) <- gsub("\\.|_", " ", colnames(df))
    df
  }

  prettyKable <- function(df, digits=2, row.names=FALSE){
  kable(prettyColNamesDf(df), digits = digits, row.names = row
.names )
  }



is.gui <- function(x){
  sapply(x, function(x){
  retVal <- FALSE
  if(length(grep("C25", x)) > 0){
    retVal <- TRUE
  }

  retVal
  })
}
getSplitWeightGUIEXT <- function(consmuptions, id, day) {
  if (nrow(consmuptions) == 0) {
    print(id)
    print(day)
    stop(consmuptions)
  }
  agg <- aggregate(FWT ~ is.gui(GUI_CODE_EXT), consmuptions, s
um)
  nonguiIdx <- which(agg$`is.gui(GUI_CODE_EXT)` == FALSE)
  guiIdx <- which(agg$`is.gui(GUI_CODE_EXT)` )
  nonGuiFWT <- NA
  if (length(nonguiIdx) > 0) {
    nonGuiFWT <- unlist(agg$FWT[nonguiIdx])
  }
  guiFWT <- NA
  if (length(guiIdx) > 0) {
    guiFWT <- unlist(agg$FWT[guiIdx])
  }

  list(gui = guiFWT, nonGui = nonGuiFWT)

}
```

```
  prettyColNamesDf <- function(df){
    colnames(df) <- gsub("\\.|_", " ", colnames(df))
    df
  }

  prettyKable <- function(df, digits=2, row.names=FALSE){
  kable(prettyColNamesDf(df), digits = digits, row.names = row
.names )
  }



foodDataGUIMapped <- read.csv("foodDataGUIMappedV2.csv")


foodDataSPSS <- read.spss(file="npns-food-file-4R.sav")

# Loaded initially to map against the 77 food categories
iunaFoodCodes<- cbind.data.frame(code=seq(1,77), description=l
evels(foodDataSPSS$IUNA_NPNS_77FG))

subjects <- unique(foodDataGUIMapped$SUBJECID)
survayDays <- unique(foodDataGUIMapped$SURVDAY)
colSelection <- c('SUBJECID', 'SURVDAY', 'MTYPE', 'TIME', 'IUN
A_NPNS_77FG', 'FWT', 'GUI_CODE')
```

The algorithm starts from the IUNA data mapped with the GUI code. There are 6 variables taken in account: 1. Subject ID 2. Meal Type 3. Time 4. IUNA NPNS 77 FG 5. Food weight 6. GUI code

The algorithm iterates through each subject and day of the survey and identifies each meal that the subject consumed by sorting the records in the diary by time of the day and meal type. All the records at the same time and with same meal are considered to be part of a meal.

Below an example of the data relate to subject 108 for the first day of the survey

```
library(knitr)

## Warning: package 'knitr' was built under R version 3.4.1

foodDataSample <- foodDataGUIMapped[foodDataGUIMapped$SUBJECID
==108 & foodDataGUIMapped$SURVDAY == 1, colSelection]
reordered <- with(foodDataSample, order(SURVDAY,TIME,MTYPE))
kable(foodDataSample[reordered,])
```

	SUBJECID	SURVDAY	MTYPE	TIME	IUNA_NPNS_77FG	FWT	GUI_CODE
1	108	1	1	07:45	57	4	NA
3	108	1	1	07:45	65	20	C25k
13	108	1	1	07:45	22	3	NA
15	108	1	1	07:45	4	30	NA
23	108	1	1	07:45	7	122	NA
2	108	1	2	12:00	65	88	C25k
4	108	1	2	12:00	37	50	C25a
5	108	1	2	12:00	35	55	NA
6	108	1	2	12:00	37	14	C25a
11	108	1	2	12:00	36	176	C25a
12	108	1	2	12:00	15	100	C25i
16	108	1	2	12:00	15	90	C25i
20	108	1	2	12:00	8	17	C25g
21	108	1	2	12:00	14	21	C25i
14	108	1	7	16:00	8	11	C25g
19	108	1	7	16:00	8	17	C25g
7	108	1	5	17:00	35	57	NA
8	108	1	5	17:00	8	3	C25g
9	108	1	5	17:00	25	31	NA
10	108	1	5	17:00	29	24	C25b
17	108	1	5	17:00	16	34	NA
18	108	1	5	17:00	43	30	NA
22	108	1	5	17:00	9	28	C25g

The algorithm identifes each meal for example the meal at 07:45 is separate from the one at 12:00

```
kable((foodDataSample[reordered,])[1:5,])
```

	SUBJECID	SURVDAY	MTYPE	TIME	IUNA_NPNS_77FG	FWT	GUI_CODE
1	108	1	1	07:45	57	4	NA
3	108	1	1	07:45	65	20	C25k
13	108	1	1	07:45	22	3	NA
15	108	1	1	07:45	4	30	NA
23	108	1	1	07:45	7	122	NA

```
kable((foodDataSample[reordered,])[6:14,])
```

	SUBJECID	SURVDAY	MTYPE	TIME	IUNA_NPNS_77FG	FWT	GUI_CODE

2	108	1	2	12:00	65	88	C25k
4	108	1	2	12:00	37	50	C25a
5	108	1	2	12:00	35	55	NA
6	108	1	2	12:00	37	14	C25a
11	108	1	2	12:00	36	176	C25a
12	108	1	2	12:00	15	100	C25i
16	108	1	2	12:00	15	90	C25i
20	108	1	2	12:00	8	17	C25g
21	108	1	2	12:00	14	21	C25i

Once each meal is identified we restructure the data a tree shaped where each branch is a subject and each leaf is a list of days and corresponding meals.

Once this tree is built we perform two distinct analysis by going through each of the leaves of the tree and keeping the metadata (subject information) of the branch:

5. We use the set of distinct GUI codes of each meal and NA when not present to build a unique identifier for the meal. We call this *key*
6. As per above but instead we use the IUNA F77 codes.

The first analysis reported the top meal by frequency, the second analysis was extended to distinguish between snacks and main meals and focus on cariogenic food and their interaction.

Using IUNA NPNS FG 77

```
bySubject <- lapply(subjects, function(s){
  byDay <- lapply(survayDays, function(d){
    dailyMeals <- foodDataGUIMapped[foodDataGUIMapped$SUBJECID
== s & foodDataGUIMapped$SURVDAY == d,]
    dailyMeals <- dailyMeals[with(dailyMeals, order(MTYPE, TIM
E)),]
    mealTime <- unique(dailyMeals[, c('MTYPE', 'TIME')])
    byMeal <- apply(mealTime, 1, function(idx){
      meal <- as.integer(idx['MTYPE'])
      time <- idx['TIME']
      foodsInMeal <- dailyMeals[dailyMeals$TIME == time & dail
yMeals$MTYPE == meal, ]
      if(nrow(foodsInMeal) == 0){
        stop(foodsInMeal)
      }
      foodsInMeal <- foodsInMeal[order(foodsInMeal$IUNA_NPNS_7
7FG), c('IUNA_NPNS_77FG', 'FWT')]
      foodsInMeal
    })
```

```
    list(meals=mealTime, foodInMeals=byMeal, SURVDAY=d)
  })
  list(id=s, diary=byDay)
})



keyedConsumption <- lapply(bySubject, function(subject){
  id <- subject$id
  diaryDays <- subject$diary
  lapply(diaryDays, function(day){
    meals <- day$meals
    foodInMeals <- day$foodInMeals
    t(sapply(1:nrow(meals), function(i){
      foods <- foodInMeals[[i]]
      key <- paste(iunaFoodCodes[unique(foods$IUNA_NPNS_77FG),
'description'], sep="-", collapse = "-")
      meal <- meals[i,1]
      list(key=key, MTYPE=meal, subject=id, fwt=sum(foods$FWT)
, SURVDAY=day$SURVDAY)
    }))
  })
})



keyedConsumptionDf <-
  as.data.frame(do.call(rbind, lapply(1:length(keyedConsumptio
n), function(i) {
    if (length(keyedConsumption[[i]]) != 4) {
      stop(keyedConsumption)
    }
    rbind(keyedConsumption[[i]][[1]],
          keyedConsumption[[i]][[2]],
          keyedConsumption[[i]][[3]],
          keyedConsumption[[i]][[4]])
  })))

for(i in 1:ncol(keyedConsumptionDf)) {
  keyedConsumptionDf[, i] <- unlist(keyedConsumptionDf[, i])
}
str(keyedConsumptionDf)

## 'data.frame':    2676 obs. of  5 variables:
##  $ key    : chr  "Wholemeal & brown bread & rolls-Other bre
akfast cerals-Other spreading fats-Sugars, syrups, preserves &
```

```
sweeten"| __truncated__ "Biscuits including crackers-Cheeses-Y
ogurts-Fruit juices-Bananas-Other fruit-Other beverages" "Bisc
uits including crackers-Cakes, pastries & buns-Ice creams-Pota
toes-Peas, beans & lentils-Fruit juices-Bacon & ham" "Biscuits
including crackers" ...
##  $ MTYPE  : int  1 2 5 7 1 2 5 7 1 2 ...
##  $ subject: int  108 108 108 108 108 108 108 108 108 108 ..
.
##  $ fwt    : int  179 611 207 28 133 569 78 294 147 476 ...
##  $ SURVDAY: int  1 1 1 1 2 2 2 2 3 3 ...

#### Report this !!! what about fruit as snack

keyFreq <-
sqldf("select key KEY, case when MTYPE in (6,7,8,11) then 'sna
ck' else 'main' end Meal_Type,
                    count(subject) Total_number_EO,
                    count(distinct subject) Total_subject_coun
t,
                    AVG(FWT) Average_FWT,
                    STDEV(FWT) SD_FWT
                    from keyedConsumptionDf
                    group by key, case when MTYPE in (6,7,8,11
) then 'snack' else 'main' end
                    order by Total_number_EO desc")

prettyKable(keyFreq[1:20,], digits = 0, row.names = TRUE)
```

	KEY	Meal Type	Total number EO	Total subject count	Average FWT	SD FWT
1	Other fruit	snack	70	43	90	59
2	RTEBC-Whole milk	main	62	36	172	71
3	Whole milk	main	42	20	190	85
4	Chocolate confectionery	snack	38	30	24	11
5	Biscuits including crackers	snack	37	26	24	12
6	Yogurts	snack	23	17	119	52
7	Non-chocolate confectionery	snack	22	20	26	22
8	Savoury snacks	snack	20	17	23	13
9	Bananas	snack	19	16	110	54
10	Biscuits including crackers-Whole milk	snack	19	8	175	59
11	Ice creams	snack	18	16	57	21
12	Other milks & milk based beverages	main	18	5	194	100
13	RTEBC-Whole milk-Sugars, syrups, preserves & sweeteners	main	18	11	188	76
14	Other beverages	main	17	12	107	70

15	Other beverages-Squashes, cordials & fruit juice drinks	main	17	10	200	61
16	RTEBC-Low fat, skimmed & fortified milks	main	17	7	176	85
17	Other fruit-Other beverages	snack	14	11	240	145
18	Citrus fruits	snack	12	6	98	70
19	Whole milk	snack	12	7	199	64
20	Yogurts	main	12	7	110	38

Cariogenic foods only

```
getSplitWeightCariogenic <- function(consmuptions, id, day){
  if(nrow(consmuptions) == 0){
    print(id)
    print(day)
    stop(consmuptions)
  }
  cgFWTVec <- rep(0, length(cariogenicIUNAFoodCode))
  agg <- aggregate(FWT~IUNA_NPNS_77FG, consmuptions, sum)
  cgIdx <- which(agg$IUNA_NPNS_77FG %in% cariogenicIUNAFoodCod
e)
  ncgIdx <- which(!(agg$IUNA_NPNS_77FG %in% cariogenicIUNAFood
Code))
  cgInMeal <- agg$IUNA_NPNS_77FG[cgIdx]
  cgIUNAIdx <- which(cariogenicIUNAFoodCode %in%  cgInMeal)
  cgFWTVec[cgIUNAIdx] <- agg$FWT[cgIdx]

  ncgFWT<-NA
  if(length(ncgIdx) > 0){
    ncgFWT <- sum(unlist(agg$FWT[ncgIdx]))
  }
 cgFWT <- NA
  if(length(cgIdx) > 0){
    cgFWT <- sum(unlist(agg$FWT[cgIdx]))
  }

  list(ncg=ncgFWT, cg=cgFWT, cgVec=cgFWTVec)

}


keyedConsumption <- lapply(bySubject, function(subject){
  id <- subject$id
  diaryDays <- subject$diary
  lapply(diaryDays, function(day){
    meals <- day$meals
    foodInMeals <- day$foodInMeals
    t(sapply(1:nrow(meals), function(i){
```

```
      foods <- foodInMeals[[i]]
      #key <- paste(iunaFoodCodes[unique(foods$IUNA_NPNS_77FG)
, 'description'], sep="-", collapse = "-")
      foodCodes <- sort(unique(foods$IUNA_NPNS_77FG))
      crgKeyPart <- cariogenicIUNAFoodCode[cariogenicIUNAFoodC
ode %in% foodCodes]
      crgKeyPartDesc <- iunaFoodCodes[crgKeyPart, 'description
']
      key <- paste(crgKeyPart, sep = "-", collapse = "-")
      if(length(crgKeyPart) != length(foodCodes)){
        key <- paste(key, 0, sep="-")
      }
      keyDesc <- paste(crgKeyPartDesc, sep = "-", collapse = "
-")
      if(length(crgKeyPart) != length(foodCodes)){
        # 2017-04-17 remapping NCG to NCF
        keyDesc <- paste(keyDesc, 'NCF', sep="-")
      }
      meal <- meals[i,1]
      weightSplit <- getSplitWeightCariogenic(foods, id, day)
      retVal <- list(key=key, MTYPE=meal, subject=id, fwt=sum(
foods$FWT), SURVDAY=day$SURVDAY,
           cgFWT=weightSplit$cg, ncgFWT=weightSplit$ncg, keyDe
sc=keyDesc)

      cgList <- as.list(weightSplit$cgVec)
      cgList <- setNames(cgList, paste("I_", cariogenicIUNAFoo
dCode , sep=""))
      retVal <- append(retVal, cgList)
      retVal

    }))
  })
})

keyedConsumptionCGDf <- as.data.frame(do.call(rbind, lapply(1:
length(keyedConsumption), function(i){
  if(length(keyedConsumption[[i]]) != 4){
    stop(keyedConsumption)
  }
  rbind(keyedConsumption[[i]][[1]], keyedConsumption[[i]][[2]]
, keyedConsumption[[i]][[3]], keyedConsumption[[i]][[4]])
})))

for(i in 1:ncol(keyedConsumptionCGDf)){
  keyedConsumptionCGDf[,i] <- unlist(keyedConsumptionCGDf[,i])
}
str(keyedConsumptionCGDf)
```

```
## 'data.frame':    2676 obs. of  21 variables:
##  $ key    : chr  "57-0" "8-35-0" "8-9-16-35-0" "8" ...
##  $ MTYPE  : int  1 2 5 7 1 2 5 7 1 2 ...
##  $ subject: int  108 108 108 108 108 108 108 108 108 108 ..
.
##  $ fwt    : int  179 611 207 28 133 569 78 294 147 476 ...
##  $ SURVDAY: int  1 1 1 1 2 2 2 2 3 3 ...
##  $ cgFWT  : int  4 72 122 28 22 9 NA NA 25 39 ...
##  $ ncgFWT : int  175 539 85 NA 111 560 78 294 122 437 ...
##  $ keyDesc: chr  "Sugars, syrups, preserves & sweeteners-NC
F" "Biscuits including crackers-Fruit juices-NCF" "Biscuits in
cluding crackers-Cakes, pastries & buns-Ice creams-Fruit juice
s-NCF" "Biscuits including crackers" ...
##  $ I_6    : num  0 0 0 0 18 0 0 0 21 0 ...
##  $ I_8    : num  0 17 3 28 0 0 0 0 0 0 ...
##  $ I_9    : num  0 0 28 0 0 0 0 0 0 0 ...
##  $ I_16   : num  0 0 34 0 0 0 0 0 0 0 ...
##  $ I_17   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ I_18   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ I_35   : num  0 55 57 0 0 5 0 0 0 20 ...
##  $ I_39   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ I_57   : num  4 0 0 0 4 4 0 0 4 0 ...
##  $ I_58   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ I_59   : num  0 0 0 0 0 0 0 0 0 19 ...
##  $ I_66   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ I_68   : num  0 0 0 0 0 0 0 0 0 0 ...

mapSnacks <- function(x){
  isSnack <- x %in% c(6,7,8,11)
  retVal <- 0 #"MainMeal"
  if(isSnack){
    retVal <- 1 #"Snack"
  }
  retVal
}

keyedConsumptionCGDf$mealType <- sapply(keyedConsumptionCGDf$M
TYPE, mapSnacks)


keyCGFreq <- sqldf(
  "

  select keyDesc, mealtype,
  COUNT(subject) subjectCount,
  SUM(fq) totalFQ,
  AVG(fq) avgFQ,
  SUM(ct_dist_meals) total_ct_dist_meals,
```

```
  AVG(avgDailyIntake) avgDailyIntake,
  AVG(avgDailyIntakeCg) avgDailyIntakeCg,
  AVG(avgDailyIntakeNcg) avgDailyIntakeNcg,
  MAX(avgFWT) maxAvgFWT,
  MIN(avgFWT) minAvgFWT,
  AVG(avgFWT) avgFWT,
  AVG(avgFWTcg) avgFWTcg,
  AVG(avgFWTncg) avgFWTncg,
  MAX(avgPcCg) maxAvgPcCg,
  MIN(avgPcCg) minAvgPcCg,
  AVG(avgPcCg) avgPcCg,
  AVG(I_6) I_6, AVG(I_8) I_8, AVG(I_9) I_9, AVG(I_16) I_16, AV
G(I_17) I_17,
  AVG(I_18) I_18, AVG(I_35) I_35, AVG(I_39) I_39, AVG(I_57) I_
57,
  AVG(I_58) I_58, AVG(I_59) I_59, AVG(I_66) I_66, AVG(I_68) I_
68
  from
  (
  select keyDesc, subject, mealType,
  COUNT(subject) fq,
  COUNT(DISTINCT mealType) ct_dist_meals,
  SUM(fwt)/4 avgDailyIntake,
  SUM(cgFWT)/4 avgDailyIntakeCg,
  SUM(ncgFWT)/4 avgDailyIntakeNcg,
  AVG(fwt) avgFWT,
  AVG(cgFWT) avgFWTcg,
  AVG(ncgFWT) avgFWTncg,
  AVG(cgFWT*1.0/fwt*100) avgPcCg,
  AVG(I_6) I_6, AVG(I_8) I_8, AVG(I_9) I_9, AVG(I_16) I_16, AV
G(I_17) I_17,
  AVG(I_18) I_18, AVG(I_35) I_35, AVG(I_39) I_39, AVG(I_57) I_
57,
  AVG(I_58) I_58, AVG(I_59) I_59, AVG(I_66) I_66, AVG(I_68) I_
68
  from keyedConsumptionCGDf
  group by keyDesc, subject, mealType
  ) t
  group by keyDesc, mealtype
  "
)

IMeals  <- sqldf("
select mealtype,
CASE
WHEN I_6 > 0 THEN  6
WHEN I_8 > 0 THEN  8
WHEN I_9 > 0 THEN  9
```

```
WHEN I_16 > 0 THEN 16
WHEN I_17 > 0 THEN 17
WHEN I_18 > 0 THEN 18
WHEN I_35 > 0 THEN 35
WHEN I_39 > 0 THEN 39
WHEN I_57 > 0 THEN 57
WHEN I_58 > 0 THEN 58
WHEN I_59 > 0 THEN 59
WHEN I_66 > 0 THEN 66
WHEN I_68 > 0 THEN 68 ELSE
0
END cgIUNACode,

                                COUNT(DISTINCT subject) subjectC
ount,
                                SUM(fq) totalFQ,
                                AVG(fq) avgFQ,
                                SUM(ct_dist_meals) total_ct_dist
_meals,
                                AVG(avgDailyIntake) avgDailyInta
ke,
                                AVG(avgDailyIntakeCg) avgDailyIn
takeCg,
                                AVG(avgDailyIntakeNcg) avgDailyI
ntakeNcg,
                                MAX(avgFWT) maxAvgFWT,
                                MIN(avgFWT) minAvgFWT,
                                AVG(avgFWT) avgFWT,
                                AVG(avgFWTcg) avgFWTcg,
                                AVG(avgFWTncg) avgFWTncg,
                                MAX(avgPcCg) maxAvgPcCg,
                                MIN(avgPcCg) minAvgPcCg,
                                AVG(avgPcCg) avgPcCg,
                                AVG(I_6) I_6, AVG(I_8) I_8, AVG(
I_9) I_9, AVG(I_16) I_16, AVG(I_17) I_17,
                                AVG(I_18) I_18,  AVG(I_35) I_35,
AVG(I_39) I_39, AVG(I_57) I_57,
                                AVG(I_58) I_58, AVG(I_59) I_59,
AVG(I_66) I_66, AVG(I_68) I_68
                                from
                                (
                                select key, subject, mealType,
                                COUNT(subject) fq,
                                COUNT(DISTINCT mealType) ct_dist
_meals,
                                SUM(fwt)/4 avgDailyIntake,
                                SUM(cgFWT)/4 avgDailyIntakeCg,
                                SUM(ncgFWT)/4 avgDailyIntakeNcg,
```

```
                              AVG(fwt) avgFWT,
                              AVG(cgFWT) avgFWTcg,
                              AVG(ncgFWT) avgFWTncg,
                              AVG(cgFWT*1.0/fwt*100) avgPcCg,
                              AVG(I_6) I_6, AVG(I_8) I_8, AVG(
I_9) I_9, AVG(I_16) I_16, AVG(I_17) I_17,
                              AVG(I_18) I_18, AVG(I_35) I_35 ,
AVG(I_39) I_39, AVG(I_57) I_57,
                              AVG(I_58) I_58, AVG(I_59) I_59,
AVG(I_66) I_66, AVG(I_68) I_68
                              from keyedConsumptionCGDf
                              group by key, subject, mealType
                              ) t
                              group by mealtype,cgIUNACode
                              ")


#write.csv(file="cariogenic-meals-aggregate.csv" , x=IMeals, r
ow.names = FALSE)

colSelection <- paste("ROUND(AVG(I_", cariogenicIUNAFoodCode,
"*1.0/fwt)*100, 0) I_" ,cariogenicIUNAFoodCode ,sep="", collap
se = ", ")


cgAggregate <- sqldf(
paste(
"
select * from (
select
  key,
  mealType,
  COUNT(subject) fq,
  COUNT(DISTINCT subject) dst,
",
colSelection,
 "
  ,
  COALESCE(ROUND(AVG(ncgFWT*1.0/fwt)*100,0), 0.0) NCF,
  keyDesc
  from keyedConsumptionCGDf
  group by key, keyDesc, mealType
 )
  where NCF < 100
  order by fq desc
  ",
sep=" "
)
```

```
)

reportCgAggregate <- cgAggregate[, c("keyDesc", "mealType", "f
q", "dst", "I_6", "I_8",  "I_16", "I_17",
                                    "I_18", "I_35", "I_39", "
I_57", "I_58", "I_59", "I_66","I_68","NCF" )]

### Table for reporting and column names
#### Change below for different column names

reportCgAggregateColNames <- c("Key Descriptor", "Meal Type",
"Frequency", "Subject count", "I_6", "I_8",  "I_16", "I_17",
                                    "I_18", "I_35", "I_39", "
I_57", "I_58", "I_59", "I_66","I_68","NCF" )

### ###

iunafoodCodes <- read.csv("IUNA-food-codes.txt", sep="=", head
er = FALSE)
coldIdx <- grep("I_", reportCgAggregateColNames)
foodCodesIx <-  as.numeric(gsub("I_" , "", reportCgAggregateCo
lNames[grep("I_", reportCgAggregateColNames)]))
reportCgAggregateColNames[coldIdx] <- as.character(iunafoodCod
es[foodCodesIx, 2])

colnames(reportCgAggregate)  <- reportCgAggregateColNames

prettyKable(reportCgAggregate, digits = 0, row.names = TRUE)
```

TOP 20 (shortened for alluvial)

```
prettyKable((reportCgAggregate[cgAggregate$NCF < 100,])[1:20,], digits = 1, row.names = TRUE)
```

Key Descriptor	Meal Type	Frequency	Subject count	RTEBC	Biscuits & crackers	Ice-creams	Desserts	Rice puddings & custard	FJ & smoothies	Tinned fruits	Sugars(SSPS)	Chocolate confection	Non-chocolate confection	Carbonated beverages	Squashes(SC&FJD)	NCF
RTEBC-NCF	0	187	85	15	0	0	0	0	0	0	0	0	0	0	0	85
Squashes, cordials & fruit juice drinks-NCF	0	172	57	0	0	0	0	0	0	0	0	0	0	0	20	80
Biscuits including crackers-NCF	1	97	53	0	17	0	0	0	0	0	0	0	0	0	0	83
Fruit juices-NCF	0	80	45	0	0	0	0	0	48	0	0	0	0	0	0	52
Squashes, cordials & fruit juice drinks-NCF	1	62	43	0	0	0	0	0	0	0	0	0	0	0	29	71

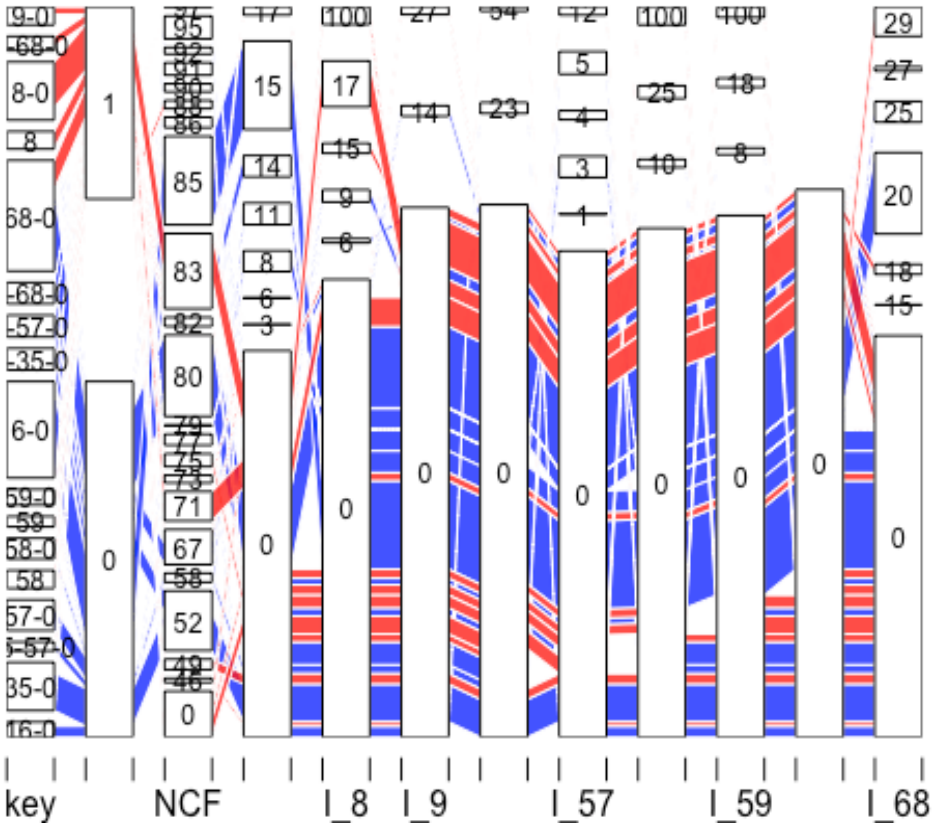
Sugars, syrups, preserves & sweeteners-NCF	0	48	29	0	0	0	0	0	0	0	5	0	0	0	0	95
RTEBC-Fruit juices-NCF	0	46	26	11	0	0	0	0	37	0	0	0	0	0	0	52
RTEBC-Sugars, syrups, preserves & sweeteners-NCF	0	45	22	14	0	0	0	0	0	0	3	0	0	0	0	83
RTEBC-Squashes, cordials & fruit juice drinks-NCF	0	44	23	8	0	0	0	0	0	0	0	0	0	0	25	67
Chocolate confectionery	1	38	30	0	0	0	0	0	0	0	0	100	0	0	0	0
Biscuits including crackers	1	37	26	0	100	0	0	0	0	0	0	0	0	0	0	0
Chocolate confectionery-NCF	1	27	22	0	0	0	0	0	0	0	0	25	0	0	0	75

Alluvials

```
library(alluvial)
```

Top 20

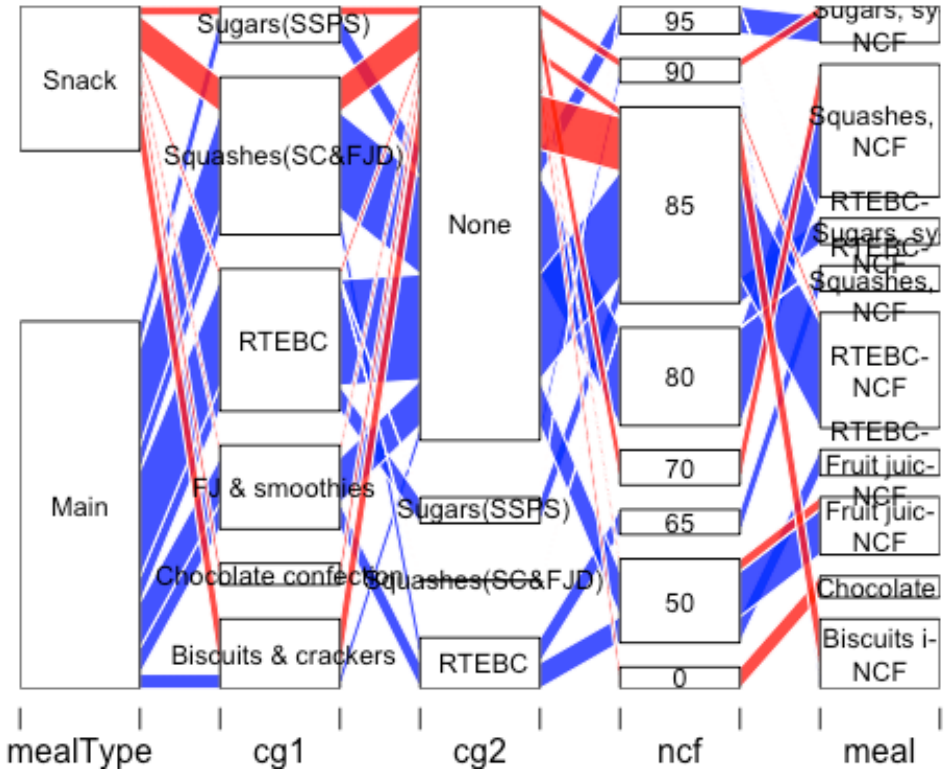
```
tmp.df <- (cgAggregate[cgAggregate$NCF < 100,])[1:20,]
tmp.df <- cgAggregate[which(cgAggregate$key %in% unique(tmp.df
$key)), ]

layout(matrix(c(1, 1), nrow = 1), widths = c(1, 1))
alluvial(
tmp.df[, c(1, 2, 18, 5:8, 13:17)],
freq = tmp.df[, 'fq'],
xw = 0,
alpha = 0.8,
gap.width = 0.25,
col = ifelse(tmp.df[, 2] == 0, "blue" , 'red'),
border = "white",
cex = 0.8,
cw = 0.3,
layer = -tmp.df$mealType
)
```



Top 10 with 2 components

```
tmp.df <- (cgAggregate[cgAggregate$NCF < 100, ])[1:10, ]
tmp.df <- cgAggregate[which(cgAggregate$key %in% unique(tmp.df
$key)), ]

#tmp.df$key <- gsub(pattern = "-0", replacement = "-NCF", x=tm
p.df$key)
tmp.df$key <- sapply(tmp.df$key, function(key){
  x <- as.numeric(unlist(strsplit(key, "-")))
  paste(ifelse(x == 0 , "NCF", substr(as.character(iunaFoodCod
es[x, 2]),1, 10)), collapse = "-\n")
}
)
tmp.df$mealType <- ifelse(tmp.df$mealType == 1, "Snack", "Main
")

# 5:17
# NCF = 18
cgIFoods <- colnames(tmp.df[, 5:17])
allVDf <- do.call("rbind", apply(tmp.df, 1,  function(row) {
  key <- row[1]
```

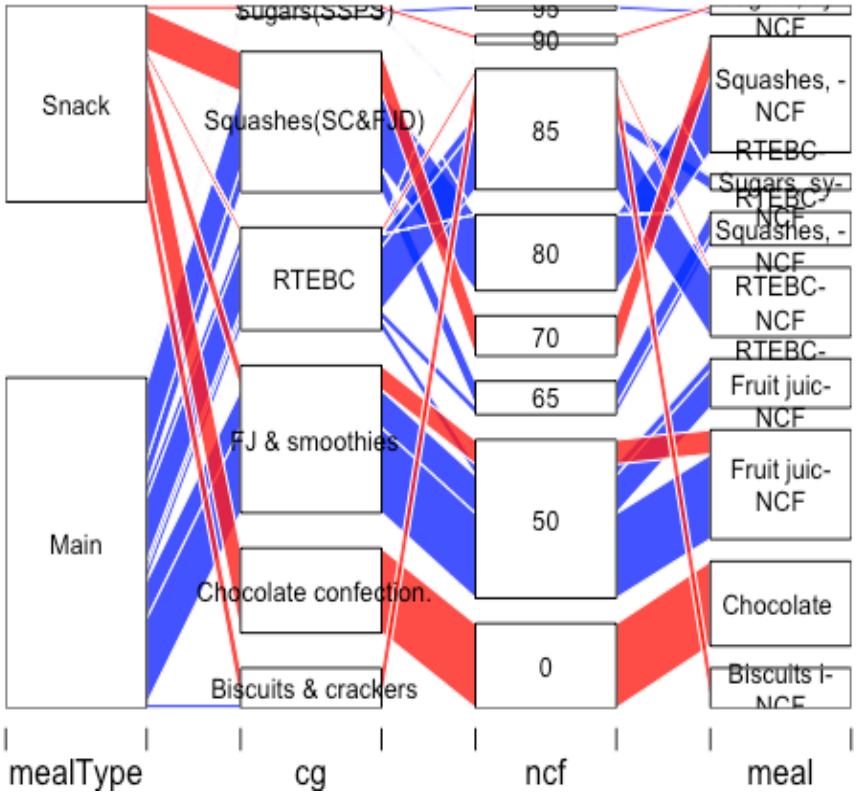
```
  mealType <- row[2]
  fq <- as.numeric(row[3])
  ncf <- as.numeric(row[18])
  ncf <- round(ncf / 5, 0 ) * 5
  iFoods <- as.numeric(row[5:17])

  iFoodsIdx <- which(iFoods > 0)
  interactingFoodsFq <- round(iFoods[iFoodsIdx] / 100 * fq, 0)
  interactingFoodsOrder <- order(interactingFoodsFq, decreasin
g = TRUE)
  interactingFoodsLabels <- cgIFoods[iFoodsIdx]
  interactingFoodsLabels <-
  as.character(iunafoodCodes[as.numeric(gsub("I_", "", interac
tingFoodsLabels)), 2])
  total <- 1 #length(interactingFoodsFq)

  retVal <- data.frame(
  fq = fq , #interactingFoodsFq,
  mealType = rep(mealType, total),
  cg1 = interactingFoodsLabels[interactingFoodsOrder[1]],
  cg2 = interactingFoodsLabels[interactingFoodsOrder[2]],
  #cg3 = interactingFoodsLabels[interactingFoodsOrder[3]],
  ncf = rep(ncf, total),
  meal = rep(key, total),
  stringsAsFactors = FALSE
  )
  retVal$cg2 <- ifelse(is.na(retVal$cg2), "None", retVal$cg2)
  #retVal$cg3 <- ifelse(is.na(retVal$cg3), "None", retVal$cg3)

  retVal
}))


alluvial(
allVDf[, -1],
freq = allVDf[, 'fq'],
xw = 0,
alpha = 0.8,
gap.width = 0.25,
col = ifelse(allVDf[, 'mealType'] == 'Main', "blue" , 'red'),
border = "white",
cex = 0.8,
cw = 0.3,
layer = -as.numeric(allVDf[, 'mealType'] == "Snack")
)
```



Top 10 1 component

```
tmp.df <- (cgAggregate[cgAggregate$NCF < 100, ])[1:10, ]
tmp.df <- cgAggregate[which(cgAggregate$key %in% unique(tmp.df
$key)), ]

tmp.df$key <- sapply(tmp.df$key, function(key){
  x <- as.numeric(unlist(strsplit(key, "-")))
  paste(ifelse(x == 0 , "NCF", substr(as.character(iunaFoodCod
es[x, 2]),1, 10)), collapse = "-\n")
}
)
tmp.df$mealType <- ifelse(tmp.df$mealType == 1, "Snack", "Main
")

# 5:17
# NCF = 18
cgIFoods <- colnames(tmp.df[, 5:17])
allVDf <- do.call("rbind", apply(tmp.df, 1,  function(row) {
```

```
  key <- row[1]
  mealType <- row[2]
  fq <- as.numeric(row[3])
  ncf <- as.numeric(row[18])
  ncf <- round(ncf / 5, 0 ) * 5
  iFoods <- as.numeric(row[5:17])

  iFoodsIdx <- which(iFoods > 0)
  interactingFoodsFq <- round(iFoods[iFoodsIdx] / 100 * fq, 0)
  interactingFoodsOrder <- order(interactingFoodsFq, decreasin
g = TRUE)
  interactingFoodsLabels <- cgIFoods[iFoodsIdx]
  interactingFoodsLabels <-
  as.character(iunafoodCodes[as.numeric(gsub("I_", "", interac
tingFoodsLabels)), 2])
  total <- length(interactingFoodsFq)

  retVal <- data.frame(
  fq = interactingFoodsFq,
  mealType = rep(mealType, total),
  cg = interactingFoodsLabels,
  ncf = rep(ncf, total),
  meal = rep(key, total),
  stringsAsFactors = FALSE
  )

  retVal
}))


alluvial(
allVDf[, -1],
freq = allVDf[, 'fq'],
xw = 0,
alpha = 0.8,
gap.width = 0.25,
col = ifelse(allVDf[, 'mealType'] == 'Main', "blue" , 'red'),
border = "white",
cex = 0.8,
cw = 0.3,
layer = -as.numeric(allVDf[, 'mealType'] == "Snack")
)
```


Snack
Main
Squashes(SC&FJD)
RTEBC
FJ & smoothies
Chocolate confection.
Biscuits & crackers
90
85
80
70
65
50
0
Squashes, -NCF
RTEBC-NCF
Fruit juic-NCF
Chocolate
mealType
cg
ncf
meal

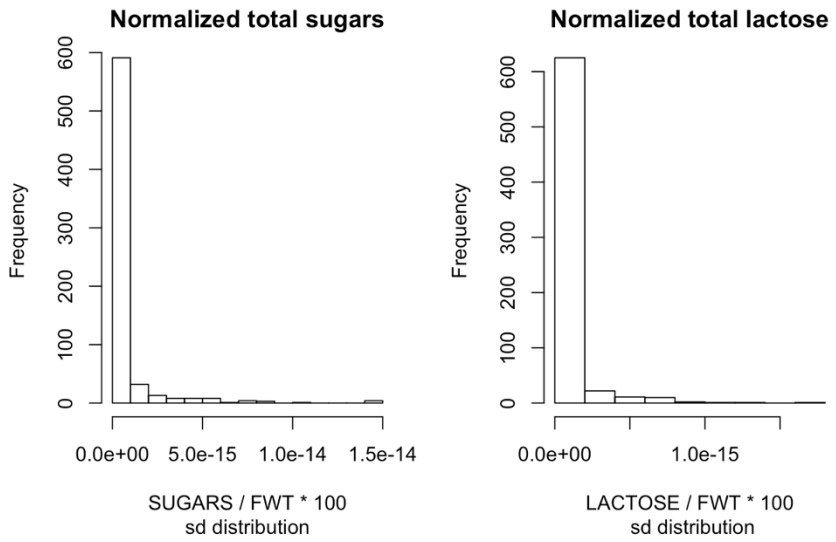
5. Free sugars mapping.Rmd

Free Sugars Mapping and Analysis

```
library(sqldf)

library(pastecs)

library(foreign)
```

This created the input for the mapping which is done in Google Sheets

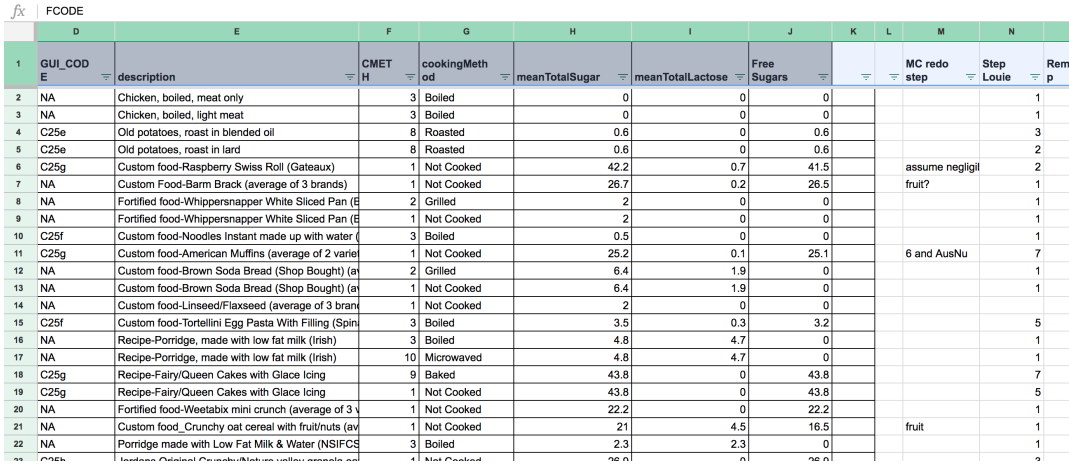
```
foodDataGUIMapped <-
  read.csv("../foodDataGUIMappedV2.csv", header = TRUE)

  foodDataGUIMapped$sugar100 <-
  foodDataGUIMapped$SUGARS / foodDataGUIMapped$FWT * 100
  foodDataGUIMapped$lactose100 <-
  foodDataGUIMapped$LACTOSE / foodDataGUIMapped$FWT * 100
  sugar100Agg <- as.matrix(aggregate(foodDataGUIMapped[,
c('sugar100', 'lactose100')], by =
  list(FCODE = foodDataGUIMapped$FCODE, CMETH =
foodDataGUIMapped$CMETH),
  function(x) {
  c(mean(x), sd(x))
  }))
  colnames(sugar100Agg) <-
  c('FCODE', 'CMETH', 'meanS100', 'sdS100', 'meanL100',
'sdL1000')
  sugar100Agg <- data.frame(sugar100Agg)

  par(mfrow = c(1, 2))
  hist(sugar100Agg$sdS100, xlab = "sd distributions",
  main = "Normalized total sugars - standard deviation
distribution")
  hist(sugar100Agg$sdL1000, xlab = "sd distribution",
  main = "Normalized total lactose - stadard deviation
distribution")
```




```
  cookingMethodsCodes <- read.table(file = "../CMETH.txt", se
p = "=")
  colnames(cookingMethodsCodes) <- c('CMETH', 'cookingMethod'
)

foods <- sqldf("
SELECT
  foods.FCODE,
  foods.IUNA_NPNS_19FG,
  foods.IUNA_NPNS_77FG,
  foods.GUI_CODE,
  foods.description,
  foods.CMETH,
  cookingMethodsCodes.cookingMethod,
  sugar100Agg.meanS100 meanTotalSugar,
  sugar100Agg.meanL100 meanTotalLactose,
  CASE
    WHEN sugar100Agg.meanS100 <= 0.0 THEN 0
    ELSE NULL
  END freeSugars
FROM (SELECT DISTINCT
  FCODE,
  IUNA_NPNS_19FG,
  IUNA_NPNS_77FG,
  Food_description_first_first description,
  CMETH,
  GUI_CODE
FROM foodDataGUIMapped) foods
LEFT JOIN sugar100Agg
  ON foods.FCODE = sugar100Agg.FCODE
  AND foods.CMETH = sugar100Agg.CMETH
LEFT JOIN cookingMethodsCodes
```

```
  ON foods.CMETH = cookingMethodsCodes.CMETH
ORDER BY sugar100Agg.meanS100 ASC, foods.description ASC, c
ookingMethodsCodes.cookingMethod ASC
      ")
```

Mapping:



FCODE

	D	E	F	G	H	I	J	K	L	M	N	
1	GUI_CODE	description	CMETH	cookingMethod	meanTotalSugar	meanTotalLactose	Free Sugars			MC redo step	Step Louie	Rem p
2	NA	Chicken, boiled, meat only	3	Boiled	0	0	0				1	
3	NA	Chicken, boiled, light meat	3	Boiled	0	0	0				1	
4	C25e	Old potatoes, roast in blended oil	8	Roasted	0.6	0	0.6				3	
5	C25e	Old potatoes, roast in lard	8	Roasted	0.6	0	0.6				2	
6	C25g	Custom food-Raspberry Swiss Roll (Gateaux)	1	Not Cooked	42.2	0.7	41.5			assume negligil	2	
7	NA	Custom Food-Barm Brack (average of 3 brands)	1	Not Cooked	26.7	0.2	26.5			fruit?	1	
8	NA	Fortified food-Whippersnapper White Sliced Pan (E	2	Grilled	2	0	0				1	
9	NA	Fortified food-Whippersnapper White Sliced Pan (E	1	Not Cooked	2	0	0				1	
10	C25f	Custom food-Noodles Instant made up with water (	3	Boiled	0.5	0	0				1	
11	C25g	Custom food-American Muffins (average of 2 varie	1	Not Cooked	25.2	0.1	25.1			6 and AusNu	7	
12	NA	Custom food-Brown Soda Bread (Shop Bought) (a	2	Grilled	6.4	1.9	0				1	
13	NA	Custom food-Brown Soda Bread (Shop Bought) (a	1	Not Cooked	6.4	1.9	0				1	
14	NA	Custom food-Linseed/Flaxseed (average of 3 bran	1	Not Cooked	2	0	0					
15	C25f	Custom food-Tortellini Egg Pasta With Filling (Spin	3	Boiled	3.5	0.3	3.2				5	
16	NA	Recipe-Porridge, made with low fat milk (Irish)	3	Boiled	4.8	4.7	0				1	
17	NA	Recipe-Porridge, made with low fat milk (Irish)	10	Microwaved	4.8	4.7	0				1	
18	C25g	Recipe-Fairy/Queen Cakes with Glace Icing	9	Baked	43.8	0	43.8				7	
19	C25g	Recipe-Fairy/Queen Cakes with Glace Icing	1	Not Cooked	43.8	0	43.8				5	
20	NA	Fortified food-Weetabix mini crunch (average of 3 v	1	Not Cooked	22.2	0	22.2				1	
21	NA	Custom food_Crunchy oat cereal with fruit/nuts (av	1	Not Cooked	21	4.5	16.5			fruit	1	
22	NA	Porridge made with Low Fat Milk & Water (NSIFCS	3	Boiled	2.3	2.3	0				1	

alt text

Then a comparison with previous Free sugar mapping (UCC) and analysis

Comparison

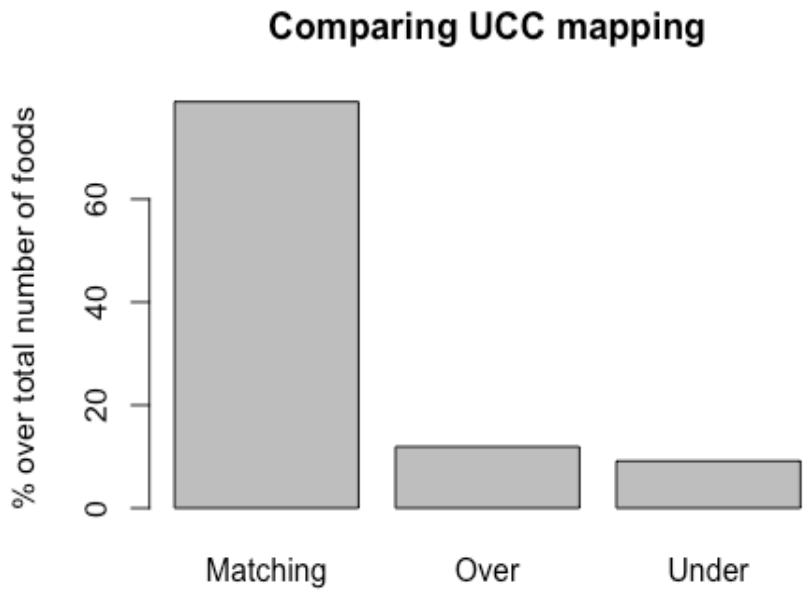
```
uccFreeSugarsList <- read.spss("UCC-free-sugars.sav")

## Warning in read.spss("UCC-free-sugars.sav"): UCC-free
-sugars.sav: Very long
## string record(s) found (record type 7, subtype 14), e
ach will be imported
## in consecutive separate variables
```

```
uccFreeSugars <- data.frame(
cbind(
FCODE = uccFreeSugarsList$FCODE,
freeSugar = uccFreeSugarsList$PHE_freesugars,
FoodDescription = uccFreeSugarsList$Food_description_first_
first,
source = uccFreeSugarsList$Source
)
)

uccFreeSugars$freeSugar <-
as.numeric(levels(uccFreeSugars$freeSugar))[uccFreeSugars$f
reeSugar]
```

```r
freeSugarMappingMC <-
read.csv("NPNS-free-sugar-mapping-MC-2017-12-24.csv", heade
r = TRUE)
freeSugarMappingMC <- unique(freeSugarMappingMC)
which(table(freeSugarMappingMC$FCODE) > 1)

## named integer(0)

foodDataGUIMappedMCMerge <-
  merge(foodDataGUIMapped, freeSugarMappingMC, by = "FCODE"
)

  dailyConsumptionAggregateSubject <- sqldf(
  "
SELECT
  FCODE FCODE_S,
  SUBJECID,
  SURVDAY,
  SUM(FWT) fwt_daily,
  SUM(SUGARS) sugars_daily,
  SUM(freeSugar/meanTotalSugar*SUGARS) free_sugars_daily,
  COUNT(SUBJECID) fq,
  COUNT(DISTINCT MEALNO) fq_meal
FROM foodDataGUIMappedMCMerge
GROUP BY FCODE, SUBJECID, SURVDAY
  "
  )

  subejctCount <- length(unique(foodDataGUIMapped$SUBJECID)
)
  # cnc -> consumer and non-consumer
  dailyConsumptionAggregate <-
  as.data.frame(as.matrix(
  aggregate(cbind(fwt_daily, fq, sugars_daily, free_sugars_
daily, fq_meal) ~
  FCODE_S, data = dailyConsumptionAggregateSubject,  functi
on(x) {
  c(
  mean = mean(x),
  quantile(x, 0.5),
  quantile(x, 0.9),
  max = max(x),
  mean.cnc = sum(x) / 4 / subejctCount
  )
  })
  ))

# cnc -> consumer and non-consumer
```

```
dailyConsumptionAggregateCNC <- sqldf(
  sprintf(
"SELECT
  FCODE FCODE_S,
  SUM(FWT) / 126.0 / 4.0 fwt_daily_nc,
  SUM(SUGARS) / %d / 4.0 sugars_daily_cnc,
  SUM(freeSugar / meanTotalSugar * SUGARS) / %d / 4.0,
  COUNT(SUBJECID) fq,
  COUNT(DISTINCT SUBJECID) consumer_count,
  COUNT(DISTINCT MEALNO) fq_meal
FROM foodDataGUIMappedMCMerge
GROUP BY FCODE",
    subejctCount,
    subejctCount
  )
)


##
comparisonDf <-
  sqldf(
"
SELECT
  a.*,
  b.freeSugar freeSugarUCC,
  b.source,
  (a.freeSugar - b.freeSugar) delta,
  CASE
    WHEN ABS(a.freeSugar - b.freeSugar) < 0.1 THEN 'Matchin
g'
    WHEN a.freeSugar > b.freeSugar THEN 'Over'
    WHEN a.freeSugar < b.freeSugar THEN 'Under'
    ELSE 'Unknown'
  END comparison,
  rs.*
FROM freeSugarMappingMC a
LEFT JOIN uccFreeSugars b
  ON b.FCODE = a.FCODE
LEFT JOIN dailyConsumptionAggregate rs
  ON a.FCODE = rs.FCODE_S
"
  )


comparisonSummary <-
  sqldf(
"SELECT
  comparison,
```

```
  COUNT(DISTINCT FCODE) count,
  MAX(ABS(delta)) max_diff
FROM comparisonDf
GROUP BY comparison"
  )

barplot(
  comparisonSummary$count / sum(comparisonSummary$count) *
100,
  names.arg = comparisonSummary$comparison,
  ylab = "% over total number of foods",
  main = "Comparing UCC mapping"
  )
```



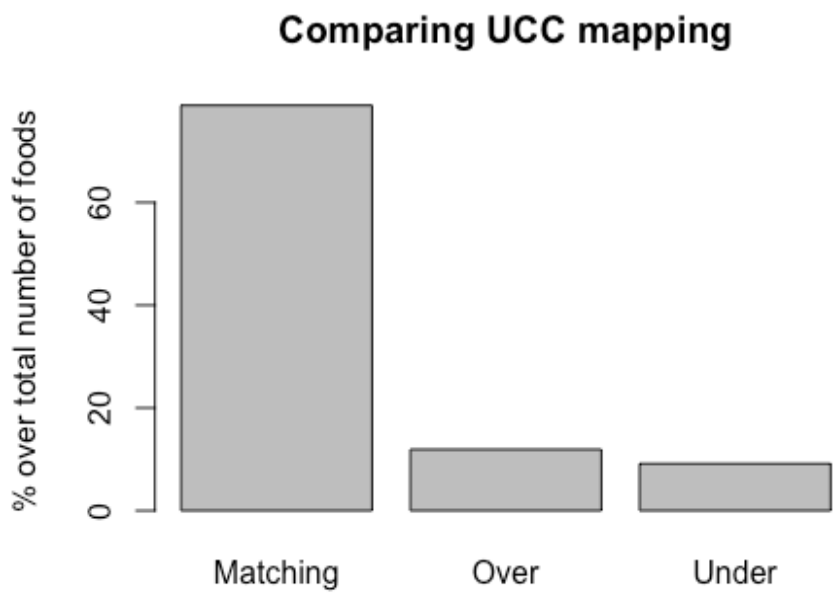

```
comparisonDf$meanTotalSugar <-
  comparisonDf$sugars_daily.mean  / comparisonDf$fwt_daily.
mean * 100

  comparisonDf$meanTotalSugar[comparisonDf$FCODE == 11590]
```

```
## [1] 23.4
```

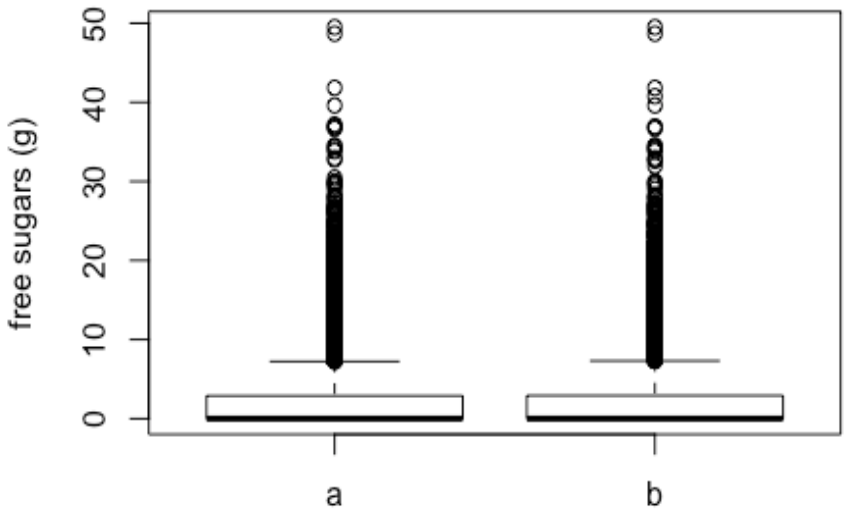
Analysis

```
foodDataGUIMappedFreeSugars <- sqldf(
"SELECT
```

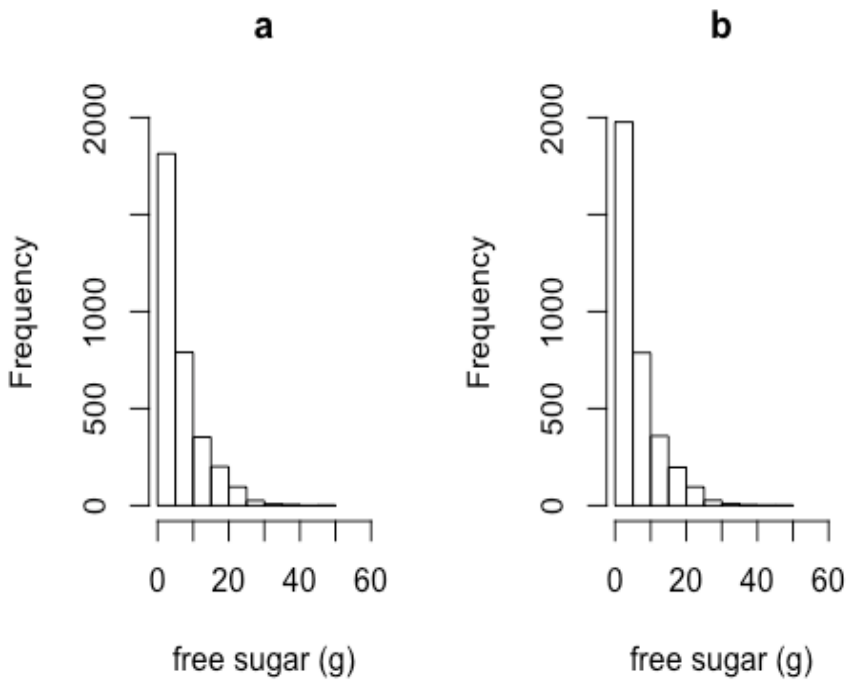
```
  f.*,
  c.meanTotalSugar,
  c.freeSugar freeSugars100,
  c.freeSugarUCC freeSugarsUCC100,
  CASE
    WHEN c.meanTotalSugar = 0 THEN 0
    ELSE f.sugars * c.freeSugar / c.meanTotalSugar
  END freeSugar,
  CASE
    WHEN c.meanTotalSugar = 0 THEN 0
    ELSE f.sugars * c.freeSugarUCC / c.meanTotalSugar
  END freeSugarUCC,
  c.comparison
FROM foodDataGUIMapped f
LEFT JOIN comparisonDf c
  ON c.FCODE = f.FCODE"
  )


  comparisonSummary <-
  sqldf(
"SELECT
  comparison,
  COUNT(DISTINCT FCODE) count
FROM foodDataGUIMappedFreeSugars
GROUP BY comparison"
  )

  barplot(
  comparisonSummary$count / sum(comparisonSummary$count) *
100,
  names.arg = comparisonSummary$comparison,
  ylab = "% over total number of foods",
  main = "Comparing UCC mapping"
  )
```




```
  ks.test(comparisonDf$freeSugar, comparisonDf$freeSugarUCC, c
orrect = TRUE)

  par(mfrow = c(1, 1))
  boxplot(
  foodDataGUIMappedFreeSugars$freeSugar,
  foodDataGUIMappedFreeSugars$freeSugarUCC,
  names = c('a', 'b'),
  ylab = "free sugars (g)"
  )
```




```
  par(mfrow = c(1, 2))
  hist(
  foodDataGUIMappedFreeSugars$freeSugar[foodDataGUIMappedFreeS
ugars$freeSugar > 0],
  xlab = "free sugar (g)",
  main = "a",
  xlim = c(0, 60),
  ylim = c(0, 2000)
  )
  hist(
  foodDataGUIMappedFreeSugars$freeSugarUCC[foodDataGUIMappedFr
eeSugars$freeSugarUCC > 0],
  xlab = "free sugar (g)",
  main = "b",
  xlim = c(0, 60),
  ylim = c(0, 2000)
  )
```


a
b
Frequency
0
500
1000
2000
0
20
40
60
free sugar (g)

6. Non-covered sugars.Rmd

Non-covered sugars analysis

The goal of this analysis was to describe the covered and non-covered sugars in GUI. The rationale is that some key foods which contains sugars are actually not covered by the GUI mapping. We used the food mapping derived in the previous analysis and explore what food is covered and not covered by the GUI mapping in respect of IUNA / NPNS data.

Objective was to give a quantitive analysis of the non-covered sugars compared with the covered ones.

Definitions

With the term **non-covered** we indicate a consumption that can not be mapped using a GUI food code. Generally, if this means that specific food item can not be mapped to a GUI code but for some food items such as pizza and similar we need to take in account also the cooking method and the meal type in order to decide whether or not the food is mapped.

Analysis

Loading the data

Note the source file used was the `foodDataGUIMappedV2.csv` which contained the new food mapping which included the meal type.

```
library(sqldf)

library(pastecs)
library(foreign)




foodDataSPSS <- read.spss(file="npns-food-file-4R.sav")
## re-encoding from CP1252

# Loaded initially to map against the 77 food categories
iunaFoodCodes<- cbind.data.frame(code=seq(1,77), description=l
evels(foodDataSPSS$IUNA_NPNS_77FG))
foodDataGUIMapped <- read.csv("foodDataGUIMappedV2.csv")
```

```
freeSugarMappingMC <- read.csv("free-sugars/NPNS-free-sugar-mapping-MC-2017-12-24.csv", header = TRUE)
freeSugarMappingMC <- unique(freeSugarMappingMC)
#which(table(freeSugarMappingMC$FCODE) > 1)

foodDataGUIMapped <- merge(foodDataGUIMapped, freeSugarMappingMC, by="FCODE")

foodDataGUIMapped$freeSugar <- with(foodDataGUIMapped, freeSugar/meanTotalSugar*SUGARS)
```

- **Aggregation process**

Aggregated the data across:

1. Subject Id
2. Day of the week
3. Day of the survay

and we compute

4. The total number of times when a non-gui consumption occurs.
5. The total food weight for non-gui consumptions.
6. The total sugars (weight) for a non-gui consumption.
7. The total sugars for a gui consumption.
8. The total number of times when a gui consumption occurs.
9. The total sugars.
10. The total number of number of consumptions.
11. The total food weight.

We then derive the following ratios

12. Ratio of the count of non-gui over total food
13. Ratio of the food weight of non-gui over total food weight.
14. Ratio of non gui total sugars over the total sugars
15. Ratio of gui total sugars over the total sugars.

```
nonGUIConsumptions <- sqldf(
  "SELECT SUBJECID, DOW, SURVDAY,
  SUM(CASE WHEN GUI_CODE IS NULL THEN 1 ELSE 0 END) non_gui_ct,
  SUM(CASE WHEN GUI_CODE IS NULL THEN FWT ELSE 0 END) non_gui_fwt,
  SUM(CASE WHEN GUI_CODE IS NULL THEN sugars ELSE 0 END) uncovered_total_sugars,
  SUM(CASE WHEN GUI_CODE IS NULL THEN 0 ELSE sugars END) covered_total_sugars,
  SUM(CASE WHEN GUI_CODE IS NULL THEN 1 ELSE 0 END) uncovered_ct_sugars,
```

```
  SUM(CASE WHEN GUI_CODE IS NULL THEN 0 ELSE 1 END) covered_ct
_sugars,
  SUM(CASE WHEN GUI_CODE IS NULL THEN freeSugar ELSE 0 END) un
covered_total_free_sugars,
  SUM(CASE WHEN GUI_CODE IS NULL THEN 0 ELSE freeSugar END) co
vered_total_free_sugars,
  COUNT(DISTINCT CASE WHEN SUGARS > 0 THEN MEALNO ELSE NULL EN
D) sugar_meal_ct,
  COUNT(DISTINCT CASE WHEN freeSugar > 0  THEN MEALNO ELSE NUL
L END) free_sugar_meal_ct,
  SUM(sugars) total_sugars,
  SUM(freeSugar) total_free_sugars,
  COUNT(*) day_ct,
  SUM(FWT) day_fwt
  FROM foodDataGUIMapped GROUP BY SUBJECID, DOW,SURVDAY"
  )

nonGUIConsumptions$ratio_ct <-
  nonGUIConsumptions$non_gui_ct / nonGUIConsumptions$day_ct

  nonGUIConsumptions$ratio_fwt <-
  nonGUIConsumptions$non_gui_fwt / nonGUIConsumptions$day_fwt

  nonGUIConsumptions$ratio_uncovered_sugars <-
  nonGUIConsumptions$uncovered_total_sugars / nonGUIConsumptio
ns$total_sugars

  nonGUIConsumptions$ratio_sugar_food <-
  nonGUIConsumptions$total_sugars / nonGUIConsumptions$day_fwt

  nonGUIConsumptions$ratio_u_sugar_food <-
  nonGUIConsumptions$uncovered_total_sugars / nonGUIConsumptio
ns$day_fwt

  nonGUIConsumptions$ratio_c_sugar_food <-
  nonGUIConsumptions$covered_total_sugars / nonGUIConsumptions
$day_fwt

  nonGUIConsumptions$ratio_uncovered_free_sugars <-
  nonGUIConsumptions$uncovered_total_free_sugars / nonGUIConsu
mptions$total_free_sugars

  nonGUIConsumptions$ratio_free_sugar_food <-
  nonGUIConsumptions$total_free_sugars / nonGUIConsumptions$da
y_fwt

  nonGUIConsumptions$ratio_u_f_sugar_food <-
  nonGUIConsumptions$uncovered_total_free_sugars / nonGUIConsu
```

```
mptions$day_fwt

  nonGUIConsumptions$ratio_c_f_sugar_food <-
  nonGUIConsumptions$covered_total_free_sugars / nonGUIConsump
tions$day_fwt
```

Exploratory

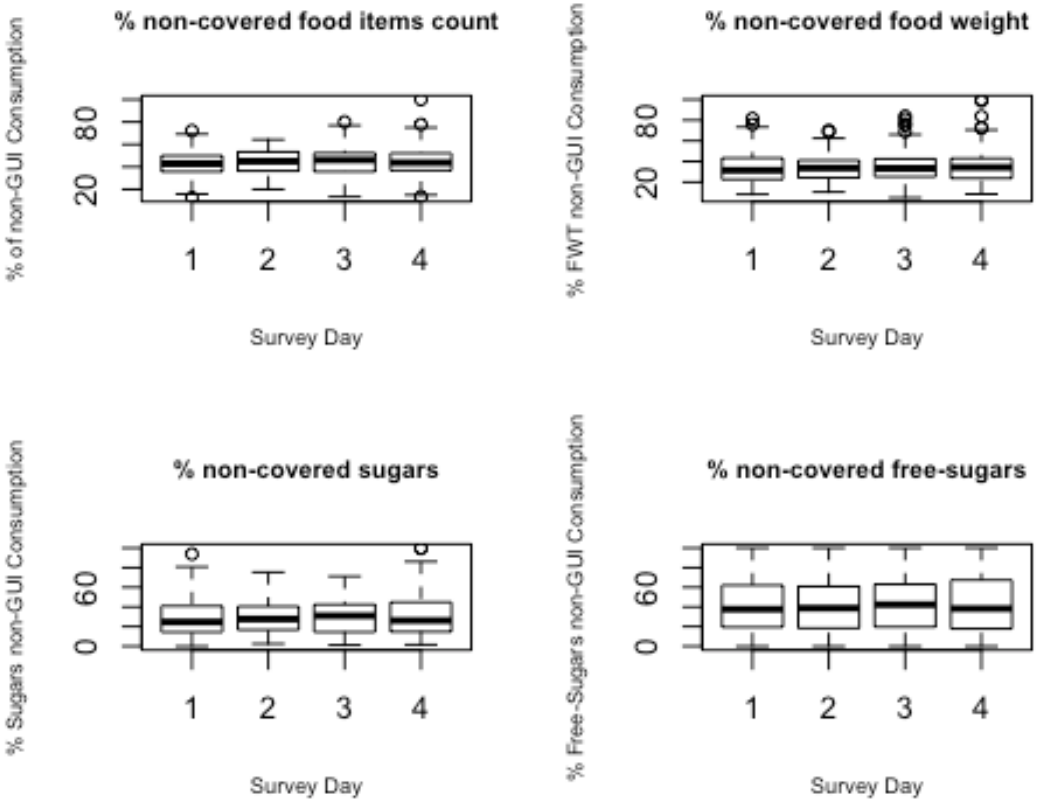
Day of the survey and Day of the week

To assess if there were noticeable differences across different days of the survey and different days of the week.

```
par(mfrow=c(2,2))
with(nonGUIConsumptions,
     list(boxplot(ratio_ct*100~SURVDAY, xlab="Survey Day", yla
b="% of non-GUI Consumption",
             main="% non-covered food items count"),
          boxplot(ratio_fwt*100~SURVDAY,  xlab="Survey Day", y
lab="% FWT non-GUI Consumption",
              main="% non-covered food weight"),
          boxplot(ratio_uncovered_sugars*100~SURVDAY,  xlab="S
urvey Day", ylab="% Sugars non-GUI Consumption",
                  main="% non-covered sugars"),
          boxplot(ratio_uncovered_free_sugars*100~SURVDAY,  xl
ab="Survey Day",
                  ylab="% Free-Sugars non-GUI Consumption",
                  main="% non-covered free-sugars")
     )
)
```

```
par(mfrow=c(2,2))
with(nonGUIConsumptions,
     list(boxplot(ratio_ct*100~DOW, xlab="Day of the week", yl
ab="% of non-GUI Consumption",
                  main="% non-covered food items count"),
          boxplot(ratio_fwt*100~DOW,  xlab="Day of the week",
ylab="% FWT non-GUI Consumption",
                  main="% non-covered food weight"),
          boxplot(ratio_uncovered_sugars*100~DOW,  xlab="Day o
f the week", ylab="% Sugars non-GUI Consumption",
                  main="% non-covered sugars"),
          boxplot(ratio_uncovered_free_sugars*100~DOW,
                  xlab="Day of the week",
                  ylab="% Free-Sugars non-GUI Consumption",
                  main="% non-covered free-sugars")
```

```
      )
)
```




```
## ratio_sugar_food
par(mfrow=c(2,2))
with(nonGUIConsumptions,
     list(boxplot(ratio_sugar_food*100, xlab="", ylab="(Sugars
/ Food Weight)%",
                  main="All foods items"),
          boxplot(ratio_c_sugar_food*100,  xlab="", ylab="(Sug
ars / Food Weight)%",
                  main="Covered"),
          boxplot(ratio_u_sugar_food*100,  xlab="", ylab="(Sug
ars / Food Weight)%",
                  main="non-covered sugar"),
          boxplot(ratio_u_f_sugar_food*100,  xlab="", ylab="(F
ree-sugars / Food Weight)%",
                  main="non-covered free-sugars")
```

Energy intake

```
enjkDF <- sqldf(
  "SELECT SUBJECID, SURVDAY,
  SUM(ENKJ) enkj
  FROM foodDataGUIMapped GROUP BY SUBJECID, SURVDAY"
  )

consumptions <-
  sqldf(
  "
  SELECT
  f.SUBJECID,
  f.SURVDAY,
  CASE WHEN GUI_CODE IS NULL THEN 'U*' ELSE GUI_CODE END GUI_C
ODE,
  SUM(sugars) sugars,
  SUM(freeSugar) free_sugars,
  COUNT(*) ct,
  SUM(FWT) fwt,
  MAX(e.enkj) enkj
  FROM foodDataGUIMapped f
  INNER JOIN enjkDF e ON e.SUBJECID = f.SUBJECID and e.SURVDAY
= f.SURVDAY
  GROUP BY f.SUBJECID, f.SURVDAY, f.GUI_CODE"
  )

  sugarsToKJ <- 17
  consumptions$enkj_sugars <-
  consumptions$sugars *  sugarsToKJ / consumptions$enkj * 100
  consumptions$enkj_free_sugars <-
  consumptions$free_sugars * sugarsToKJ / consumptions$enkj *
100


summarySubjectDayConsumptions <- aggregate(cbind(sugars, free_
sugars, enkj_sugars, enkj_free_sugars)~SUBJECID+SURVDAY, data=
consumptions, FUN=sum)

summarySubjectConsumptions <- aggregate(cbind(sugars, free_sug
ars, enkj_sugars, enkj_free_sugars)~SUBJECID, data=summarySubj
ectDayConsumptions, FUN=mean)

summaryConsumptions <- merge(summarySubjectConsumptions,  as.d
ata.frame(aggregate(enkj~SUBJECID, data=enjkDF, mean)), by="SU
BJECID")
```

```
reportingCols <-   c(
  "sugars",
  "enkj_sugars",
  "free_sugars",
  "enkj_free_sugars",
  "enkj"
  )

knitr::kable(
  pastecs::stat.desc(summaryConsumptions[, reportingCols]),
  col.names = c(
  "sugars",
  "pc_enkj_sugars",
  "free_sugars",
  "pc_enkj_free_sugars",
  "energy KJ"
  ),
  caption = "Consumer only", digits = 2)
```

Consumer only

	sugars	pc_enkj_sugars	free_sugars	pc_enkj_free_sugars	energy KJ
nbr.val	126.00	126.00	126.00	126.00	126.00
nbr.null	0.00	0.00	0.00	0.00	0.00
nbr.na	0.00	0.00	0.00	0.00	0.00
min	34.88	12.87	7.17	3.10	3045.91
max	126.02	39.30	77.38	32.59	7358.55
range	91.14	26.43	70.21	29.49	4312.65
sum	9553.01	3388.05	5045.15	1776.56	608331.15
median	74.34	26.91	37.14	13.47	4815.85
mean	75.82	26.89	40.04	14.10	4828.02
SE.mean	1.81	0.52	1.49	0.52	79.32
CI.mean.0.95	3.59	1.03	2.94	1.02	156.98
var	413.56	34.46	277.92	33.71	792755.56
std.dev	20.34	5.87	16.67	5.81	890.37
coef.var	0.27	0.22	0.42	0.41	0.18

```
knitr::kable(
pastecs::stat.desc(
 summaryConsumptions[summaryConsumptions$enkj_free_sugars >= 1
0, ]
),
digits = 1,
caption =  "> 10%"
)
```

> 10% TEI

	SUBJECID	sugars	free_sugars	enkj_sugars	enkj_free_sugars	enkj
nbr.val	94.0	94.0	94.0	94.0	94.0	94.0
nbr.null	0.0	0.0	0.0	0.0	0.0	0.0
nbr.na	0.0	0.0	0.0	0.0	0.0	0.0
min	108.0	44.6	21.6	15.6	10.1	3045.9
max	1584.0	126.0	77.4	39.3	32.6	7358.6
range	1476.0	81.4	55.8	23.7	22.4	4312.6
sum	70750.0	7566.6	4391.1	2669.3	1544.9	457252.3
median	437.5	78.5	46.2	28.2	15.6	4807.3
mean	752.7	80.5	46.7	28.4	16.4	4864.4
SE.mean	53.0	2.0	1.4	0.6	0.5	96.2
CI.mean.0.95	105.2	4.0	2.8	1.1	1.0	191.1
var	263728.3	384.5	181.2	30.6	22.4	870352.5
std.dev	513.5	19.6	13.5	5.5	4.7	932.9
coef.var	0.7	0.2	0.3	0.2	0.3	0.2

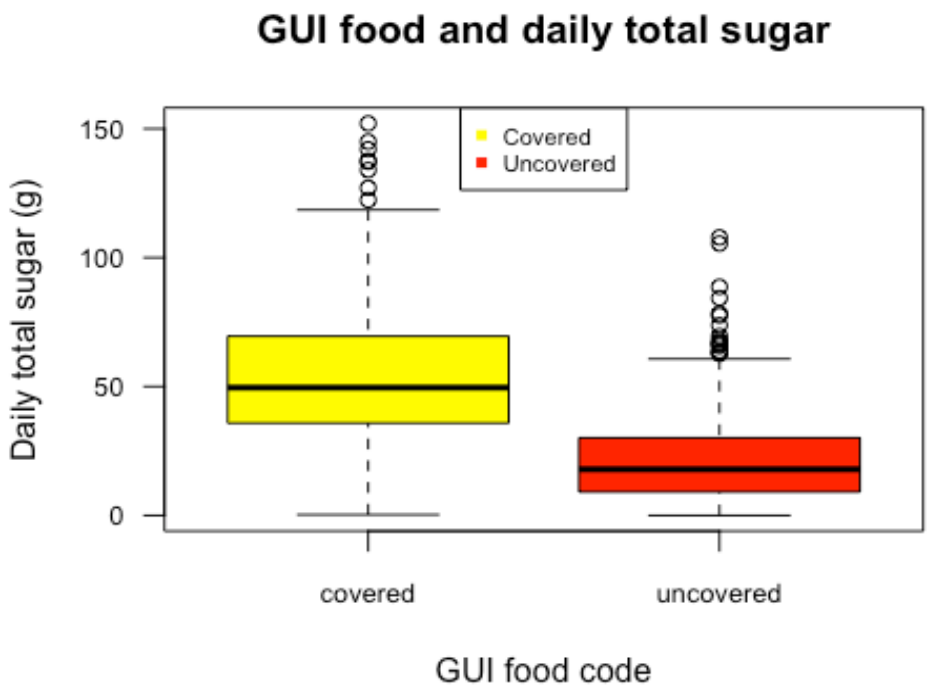
```
knitr::kable(
pastecs::stat.desc(
 summaryConsumptions[summaryConsumptions$enkj_free_sugars >= 5
, ]
),
digits = 1,
caption =  "> 5%"
)
```

> 5% TEI

	SUBJECID	sugars	free_sugars	enkj_sugars	enkj_free_sugars	enkj
nbr.val	122.0	122.0	122.0	122.0	122.0	122.0
nbr.null	0.0	0.0	0.0	0.0	0.0	0.0
nbr.na	0.0	0.0	0.0	0.0	0.0	0.0
min	108.0	34.9	10.6	12.9	5.1	3045.9
max	1584.0	126.0	77.4	39.3	32.6	7358.6
range	1476.0	91.1	66.8	26.4	27.5	4312.6
sum	94717.0	9306.6	5001.5	3290.4	1760.1	590909.9
median	449.0	75.0	38.1	27.0	13.7	4831.7
mean	776.4	76.3	41.0	27.0	14.4	4843.5
SE.mean	46.5	1.8	1.5	0.5	0.5	79.4
CI.mean.0.95	92.0	3.6	2.9	1.1	1.0	157.1
var	263723.8	414.2	257.7	35.1	31.4	768658.7
std.dev	513.5	20.4	16.1	5.9	5.6	876.7
coef.var	0.7	0.3	0.4	0.2	0.4	0.2

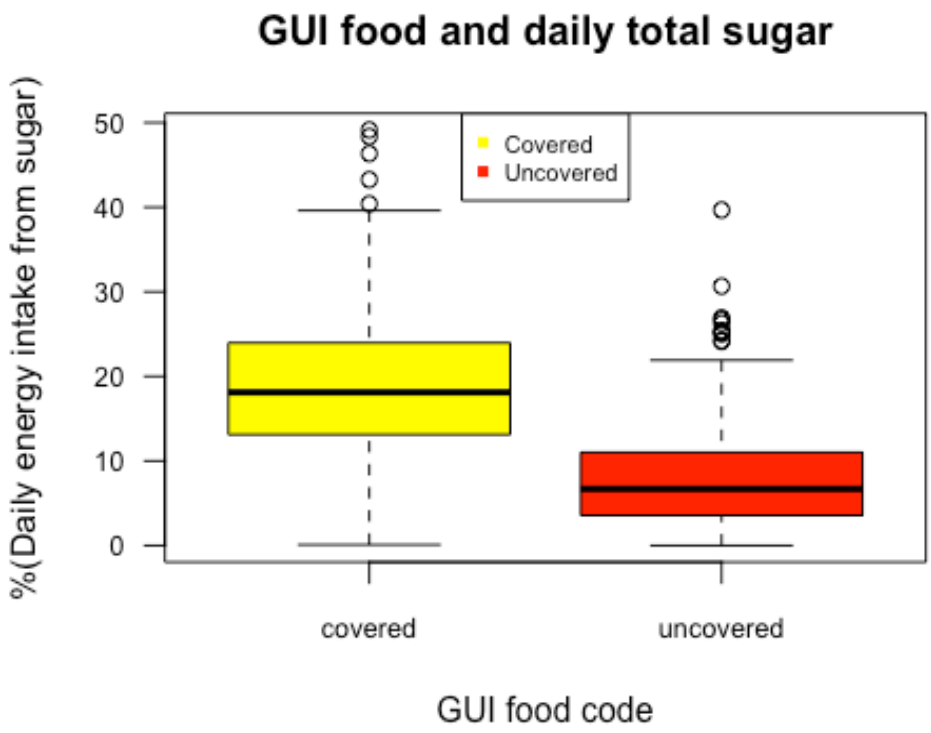
```
## Redefine consumptions
consumptions <- sqldf(
"SELECT
f.SUBJECID,
f.SURVDAY,
CASE WHEN GUI_CODE IS NOT NULL THEN 'covered'
  ELSE 'uncovered' END GUI_CODE,
 SUM(sugars) sugars,
 SUM(freeSugar) free_sugars,
 COUNT(*) ct,
 SUM(FWT) fwt,
 MAX(e.enkj) enkj
FROM foodDataGUIMapped f
INNER JOIN enjkDF e
ON e.SUBJECID = f.SUBJECID and e.SURVDAY = f.SURVDAY
GROUP BY 1, 2, 3")

consumptions$enkj_sugars <- consumptions$sugars * 17 / consump
tions$enkj * 100
consumptions$enkj_free_sugars <- consumptions$free_sugars * 17
/ consumptions$enkj * 100

#par(mfrow = c(2, 2))

with(
consumptions,
boxplot(
consumptions$sugars ~ consumptions$GUI_CODE,
xlab = "GUI food code",
ylab = "Daily total sugar (g)",
main = "GUI food and daily total sugar",
cex.axis = 0.8,
las = 1,
col = c( 'yellow', 'red')
)
)
legend(
"top",
legend = c(
"Covered",
"Uncovered"
),
col = c( 'yellow', 'red'),
pch = 15,
bg = "transparent",
```

```
cex = 0.7
)
```



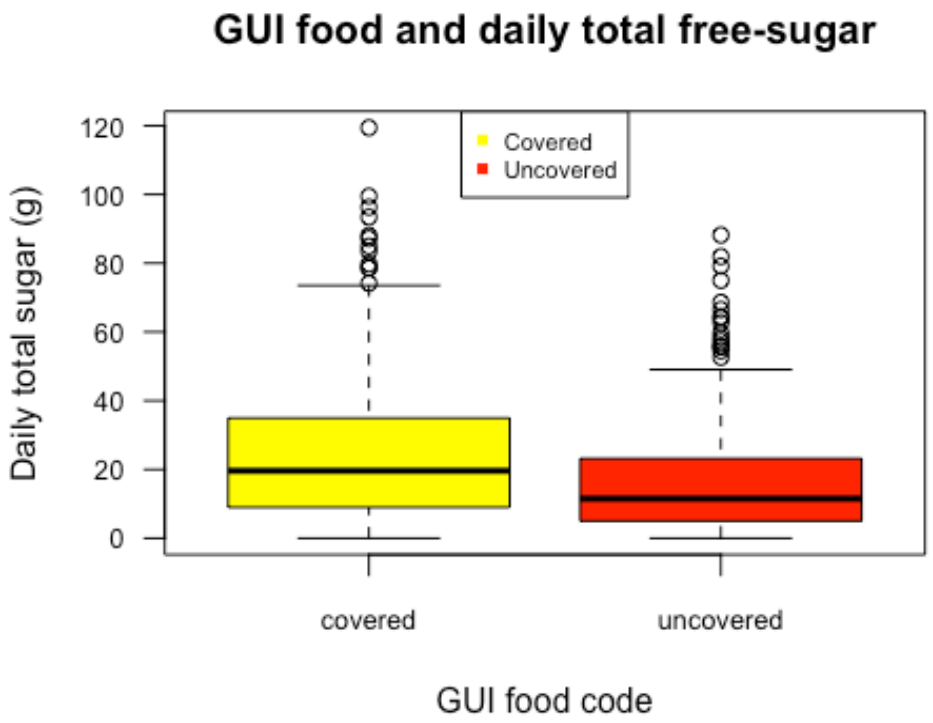

```
with(
consumptions,
boxplot(
enkj_sugars ~ GUI_CODE,
xlab = "GUI food code",
ylab = "%(Daily energy intake from sugar)",
main = "GUI food and daily total sugar",
cex.axis = 0.8,
las = 1,
col = c( 'yellow', 'red')
)
)
legend(
"top",
legend = c(
"Covered",
"Uncovered"
),
col = c( 'yellow', 'red'),
pch = 15,
bg = "transparent",
```

```
cex = 0.7
)
```



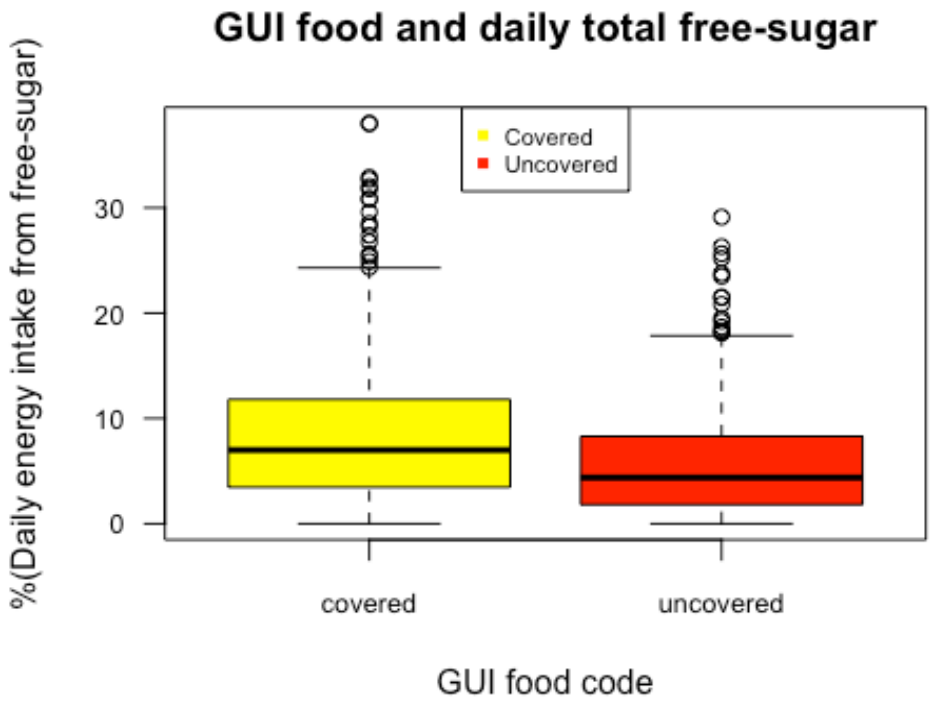

```
##
with(
consumptions,
boxplot(
consumptions$free_sugars ~ consumptions$GUI_CODE,
xlab = "GUI food code",
ylab = "Daily total sugar (g)",
main = "GUI food and daily total free-sugar",
cex.axis = 0.8,
las = 1,
col = c( 'yellow', 'red')
)
)
legend(
"top",
legend = c(
"Covered",
"Uncovered"
),
col = c( 'yellow', 'red'),
```

```
pch = 15,
bg = "transparent",
cex = 0.7
)
```



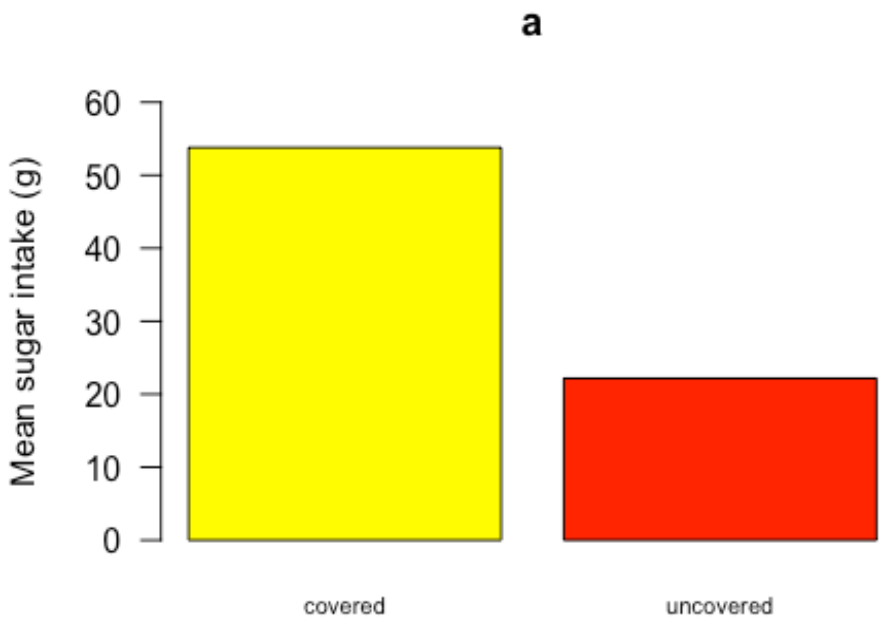

```
with(
consumptions,
boxplot(
enkj_free_sugars ~ GUI_CODE,
xlab = "GUI food code",
ylab = "%(Daily energy intake from free-sugar)",
main = "GUI food and daily total free-sugar",
cex.axis = 0.8,
las = 1,
col = c( 'yellow', 'red')
)
)
legend(
"top",
legend = c(
"Covered",
"Uncovered"
),
col = c( 'yellow', 'red'),
```

```
pch = 15,
bg = "transparent",
cex = 0.7
)
```

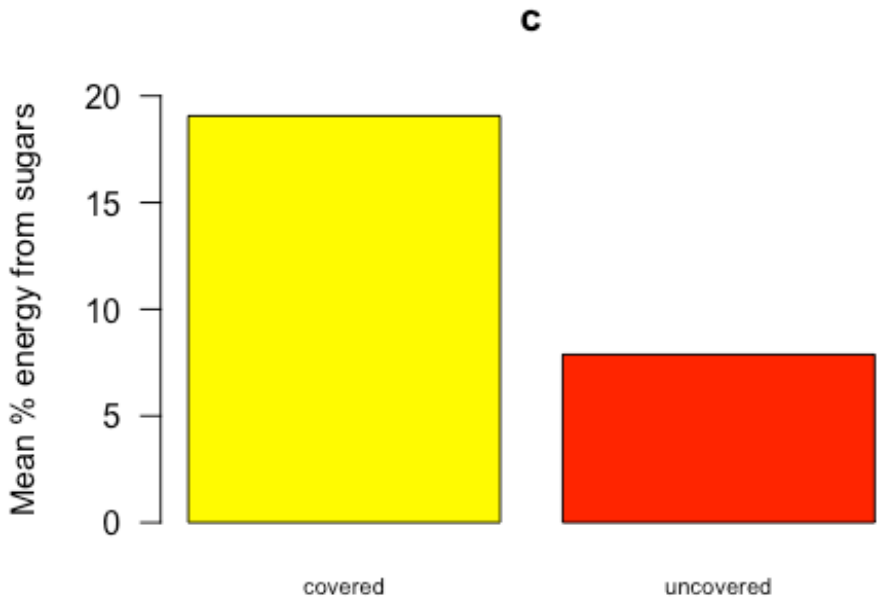



```
meanConsumptionByGUICode <- aggregate(cbind(sugars, free_sugar
s, enkj_sugars, enkj_free_sugars)~GUI_CODE, data=consumptions,
FUN=mean)

summaryConsumptions <- aggregate(cbind(sugars, free_sugars, en
kj_sugars, enkj_free_sugars)~SUBJECID+SURVDAY, data=consumptio
ns, FUN=sum)



#par(mfrow = c(2, 2))
barplot(
meanConsumptionByGUICode$sugars,
names.arg = meanConsumptionByGUICode$GUI_CODE,
cex.names = 0.7,
las = 1,
ylim=c(0,60),
ylab = "Mean sugar intake (g)",
main = "a",
```

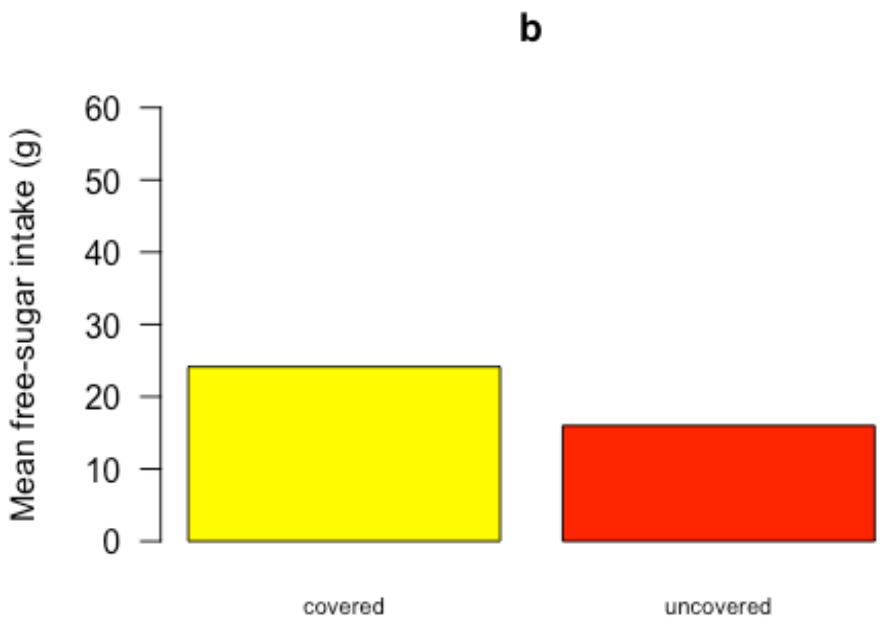
```
col = c('yellow', 'red')
)
```



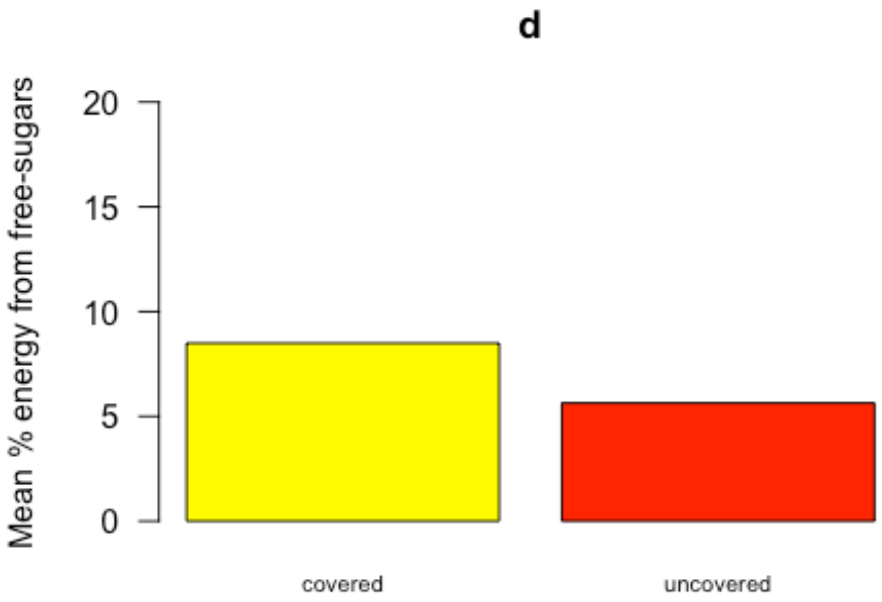

```
barplot(
meanConsumptionByGUICode$enkj_sugars,
names.arg = meanConsumptionByGUICode$GUI_CODE,
cex.names = 0.7,
las = 1,
ylim=c(0,20),
ylab = "Mean % energy from sugars",
main = "c",
col = c('yellow', 'red')
)
```




```
barplot(
meanConsumptionByGUICode$free_sugars,
names.arg = meanConsumptionByGUICode$GUI_CODE,
cex.names = 0.7,
las = 1,
ylim=c(0,60),
ylab = "Mean free-sugar intake (g)",
main = "b",
col = c('yellow', 'red')
)
```




```
barplot(
meanConsumptionByGUICode$enkj_free_sugars,
names.arg = meanConsumptionByGUICode$GUI_CODE,
cex.names = 0.7,
las = 1,
ylim=c(0,20),
ylab = "Mean % energy from free-sugars",
main = "d",
col = c('yellow', 'red')
)
```


d
Mean % energy from free-sugars
20
15
10
5
0
covered
uncovered

NPNS LEVEL

```
iunaF77MappedFoods <- read.csv("uncovered-iuna-food-code-mappi
ng.csv", stringsAsFactors = FALSE)

consumptionsNPNS <-
  sqldf(
  "
  SELECT
  f.SUBJECID,
  f.SURVDAY,
  m.category,
  SUM(sugars) sugars,
  SUM(freeSugar) free_sugars,
  COUNT(*) ct,
  SUM(FWT) fwt,
  MAX(e.enkj) enkj
  FROM foodDataGUIMapped f
  INNER JOIN enjkDF e ON e.SUBJECID = f.SUBJECID and e.SURVDAY
= f.SURVDAY
  INNER JOIN iunaF77MappedFoods m ON f.IUNA_NPNS_77FG = m.code
  GROUP BY f.SUBJECID, f.SURVDAY, m.category"
  )

  sugarsToKJ <- 17
  consumptionsNPNS$enkj_sugars <-
  consumptionsNPNS$sugars *  sugarsToKJ / consumptionsNPNS$enk
j * 100
  consumptionsNPNS$enkj_free_sugars <-
  consumptionsNPNS$free_sugars * sugarsToKJ / consumptionsNPNS
$enkj * 100


summaryNPNSSubDayConsumptions <- aggregate(cbind(sugars, free_
sugars, enkj_sugars, enkj_free_sugars, ct)~SUBJECID+SURVDAY+ca
tegory, data=consumptionsNPNS, FUN=sum)

summaryConsumptions <- aggregate(cbind(sugars, free_sugars, en
kj_sugars, enkj_free_sugars,ct)~SUBJECID+category, data=summar
yNPNSSubDayConsumptions, FUN=mean)

summaryConsumptions <- merge(summaryConsumptions,  as.data.fra
me(aggregate(enkj~SUBJECID, data=enjkDF, mean)), by="SUBJECID"
)


reportingCols <-   c(
  "sugars",
```

```
  "enkj_sugars",
  "free_sugars",
  "enkj_free_sugars",
  "enkj",
  "ct"
  )

summaryConsumptionsCategory <- aggregate( .~category ,data=sum
maryConsumptions[, -1],  function(x){
  descStats <- pastecs::stat.desc(x, basic = TRUE)
  descStats[c('min', 'max', 'mean', 'std.dev')]
})

pcConsumer <- aggregate( .~category ,data=summaryConsumptions[
, 2:3],  function(x){
  length(x)/126*100
})


summaryConsumptionsCategoryDf <- data.frame(t(summaryConsumpti
onsCategory[, -1]))
colnames(summaryConsumptionsCategoryDf) <- summaryConsumptions
Category[,1]
summaryConsumptionsCategoryDf <- summaryConsumptionsCategoryDf
[, order(colnames(summaryConsumptionsCategoryDf))]
summaryConsumptionsCategoryDf <- rbind(pcConsumer=pcConsumer[,
2], summaryConsumptionsCategoryDf)



consumerPxDf <- sqldf(
  "
  SELECT
  m.category,
  ROUND(SUM(CASE WHEN f.mtype in (6,7,8,11) THEN 1.0 ELSE 0 EN
D ) / COUNT(*) * 100 , 0) p_consumed_as_snack,
  ROUND(SUM(CASE WHEN f.mtype not in (6,7,8,11) THEN 1.0 ELSE
0 END ) / COUNT(*) * 100, 0) p_consumed_as_meal,
  COUNT(*) ct
  FROM foodDataGUIMapped f
  INNER JOIN enjkDF e ON e.SUBJECID = f.SUBJECID and e.SURVDAY
= f.SURVDAY
  INNER JOIN iunaF77MappedFoods m ON f.IUNA_NPNS_77FG = m.code
  GROUP BY m.category
  HAVING SUM (sugars) > 0
  ORDER BY m.category
  "
  )
```

```
if( length(which(consumerPxDf[,1] == colnames(summaryConsumpti
onsCategoryDf))) == length(consumerPxDf[,1]) ){
  summaryConsumptionsCategoryDf <- rbind(summaryConsumptionsCa
tegoryDf, pConsumedAsSnack=consumerPxDf[,2], pConsumedAsMeal=c
onsumerPxDf[,3] )
}



idxOrdering <- grep("^free_sugars.mean$", rownames(summaryCons
umptionsCategoryDf), )

knitr::kable(

  summaryConsumptionsCategoryDf[,order(summaryConsumptionsCate
goryDf[idxOrdering, ], decreasing = TRUE)],
  caption = "Consumer only", digits =
```

	Fruit juices and smoothies	Dairy	Soft Drinks (Non Diet)	Chocolate confectionery	Cakes and Biscuits	Non-chocolate confectionery	Sugar and syrups	Desserts and Puddings	RTEBC	Other	Bread Cereals	Fruit and vegetables
Consumers (%)	73.0	100.0	71.4	59.5	89.7	45.2	56.3	25.4	92.1	100.0	100.0	100.0
Total sugars.mean	8.7	22.0	4.8	3.6	4.7	2.2	2.6	1.2	3.2	3.0	2.5	17.3
Total sugars.SD	8.7	11.1	8.8	4.8	4.6	2.9	4.6	4.3	2.9	2.2	2.3	11.5
Free_sugars.min	0.5	0.0	0.0	0.6	0.0	1.2	0.0	0.0	0.2	0.0	0.0	0.0
Free_sugars.max	36.8	26.5	51.5	18.9	28.3	15.2	18.0	11.4	19.9	9.3	15.6	5.4
Free_sugars.mean.	8.4	8.2	4.8	3.1	4.4	2.2	2.5	0.9	3.1	1.3	0.7	0.4
Free_sugars SD	8.7	6.2	8.8	4.1	4.4	2.9	4.4	3.1	2.8	1.5	1.7	0.7
% TEI sugars.mean.	3.1	7.6	1.8	1.2	1.6	0.8	0.9	0.4	1.2	1.1	0.9	6.3
% TEI sugars.SD	3.1	3.3	3.6	1.6	1.4	1.2	1.4	1.3	1.1	0.8	0.8	4.0
%TEI free sugars.mean	3.0	2.8	1.8	1.1	1.5	0.8	0.9	0.3	1.1	0.5	0.2	0.1
% TEI free sugars.SD	3.1	1.9	3.6	1.4	1.3	1.2	1.3	0.9	1.0	0.5	0.6	0.3
Freq.mean.	0.7	3.2	0.9	0.3	0.8	0.2	0.4	0.1	0.9	5.8	1.8	3.2
Freq.std.dev	0.6	1.3	0.9	0.3	0.8	0.3	0.5	0.3	0.5	1.7	0.7	1.7
Prob. as Snack	27.0	26.0	30.0	73.0	69.0	66.0	18.0	37.0	5.0	21.0	19.0	33.0
Prob.as Meal	73.0	74.0	70.0	27.0	31.0	34.0	82.0	63.0	95.0	79.0	81.0	67.0

Simulating removal of selected cariogenic foods on %FS cut-offs

```
enjkDF <- sqldf(
  "SELECT SUBJECID, SURVDAY,
  SUM(ENKJ) enkj
  FROM foodDataGUIMapped GROUP BY SUBJECID, SURVDAY"
  )

consumptions <-
  sqldf(
  "
  SELECT
  f.SUBJECID,
  f.SURVDAY,
  SUM(sugars) sugars,
  SUM(freeSugar) free_sugars,
  COUNT(*) ct,
  SUM(FWT) fwt,
  MAX(e.enkj) enkj
  FROM foodDataGUIMapped f
  INNER JOIN enjkDF e ON e.SUBJECID = f.SUBJECID and e.SURVDAY
= f.SURVDAY
  WHERE ( CASE WHEN (f.IUNA_NPNS_77FG IN (58, 59, 57, 66, 68,
9, 8) AND f.MTYPE IN (6,7,8,11) )
          THEN 1 ELSE 0 END) = 0
  GROUP BY f.SUBJECID, f.SURVDAY"
  )

  sugarsToKJ <- 17
  consumptions$enkj_sugars <-
  consumptions$sugars *  sugarsToKJ / consumptions$enkj * 100
  consumptions$enkj_free_sugars <-
  consumptions$free_sugars * sugarsToKJ / consumptions$enkj *
100


summarySubjectDayConsumptions <- aggregate(cbind(sugars, free_
sugars, enkj_sugars, enkj_free_sugars)~SUBJECID+SURVDAY, data=
consumptions, FUN=sum)

summarySubjectConsumptions <- aggregate(cbind(sugars, free_sug
ars, enkj_sugars, enkj_free_sugars)~SUBJECID, data=summarySubj
ectDayConsumptions, FUN=mean)

summaryConsumptions <- merge(summarySubjectConsumptions,  as.d
ata.frame(aggregate(enkj~SUBJECID, data=enjkDF, mean)), by="SU
```

```
BJECID")
# Consumer and Non Consumer
# totalSubject <- length(unique(foodDataGUIMapped$SUBJECID))
# summaryConsumptionsCNC <-
#   as.data.frame(with(
#   consumptions,
#   cbind(
#   avg_sugar = sum(sugars, na.rm = TRUE) / totalSubject / 4,
#   avg_free_sugar = sum(free_sugars, na.rm = TRUE) /
#   totalSubject / 4,
#   avg_pc_energy_sugar = 100 * sum(sugars, na.rm = TRUE) * su
garsToKJ / sum(enkj, na.rm = TRUE),
#   avg_pc_energy_free_sugar =
#   100 * sum(free_sugars, na.rm = TRUE) * sugarsToKJ / sum(en
kj, na.rm = TRUE)
#   )
#   ))

# knitr::kable(summaryConsumptionsCNC, caption = "consumer and
non consumers", digits = 2)


reportingCols <-   c(
  "sugars",
  "enkj_sugars",
  "free_sugars",
  "enkj_free_sugars",
  "enkj"
  )

knitr::kable(
  pastecs::stat.desc(summaryConsumptions[, reportingCols]),
  col.names = c(
  "sugars",
  "pc_enkj_sugars",
  "free_sugars",
  "pc_enkj_free_sugars",
  "energy KJ"
  ),
  caption = "Consumer only", digits = 2)
```

	sugars	pc_enkj_sugars	free_sugars	pc_enkj_free_sugars	energy KJ
nbr.val	126.00	126.00	126.00	126.00	126.00
nbr.null	0.00	0.00	0.00	0.00	0.00
nbr.na	0.00	0.00	0.00	0.00	0.00

min	26.82	12.09	5.88	2.36	3045.91
max	121.50	37.62	62.71	22.07	7358.55
range	94.68	25.53	56.83	19.71	4312.65
sum	8367.26	2971.76	3928.26	1383.32	608331.15
median	64.55	22.91	29.06	10.73	4815.85
mean	66.41	23.59	31.18	10.98	4828.02
SE.mean	1.69	0.49	1.22	0.41	79.32
CI.mean.0.95	3.35	0.96	2.41	0.82	156.98
var	361.27	29.90	186.20	21.65	792755.56
std.dev	19.01	5.47	13.65	4.65	890.37
coef.var	0.29	0.23	0.44	0.42	0.18

```
knitr::kable(
pastecs::stat.desc(
 summaryConsumptions[summaryConsumptions$enkj_free_sugars >= 10
,                                                             ]
),
digits                          =                             1,
caption              =                        ">              10%"
)
```

> 10%	SUBJECID	sugars	free_sugars	enkj_sugars	enkj_free_sugars	enkj
nbr.val	68.0	68.0	68.0	68.0	68.0	68.0
nbr.null	0.0	0.0	0.0	0.0	0.0	0.0
nbr.na	0.0	0.0	0.0	0.0	0.0	0.0
min	108.0	41.3	22.4	15.9	10.2	3045.9
max	1584.0	121.5	62.7	37.6	22.1	7208.8
range	1476.0	80.2	40.3	21.7	11.9	4162.9
sum	53626.0	4977.9	2762.2	1779.4	984.5	327024.6
median	453.5	71.5	40.3	26.0	13.6	4785.8
mean	788.6	73.2	40.6	26.2	14.5	4809.2
SE.mean	63.9	2.3	1.2	0.6	0.4	106.2
CI.mean.0.95	127.6	4.6	2.4	1.3	0.8	212.0
var	277920.7	364.8	99.4	27.0	10.1	767398.0
std.dev	527.2	19.1	10.0	5.2	3.2	876.0
coef.var	0.7	0.3	0.2	0.2	0.2	0.2

```
knitr::kable(
pastecs::stat.desc(
 summaryConsumptions[summaryConsumptions$enkj_free_sugars >= 5,
]
```

```
),
digits                    =                    1,
caption          =                ">          5%"
)
```

> 5%	SUBJECID	sugars	free_sugars	enkj_sugars	enkj_free_sugars	enkj
nbr.val	115.0	115.0	115.0	115.0	115.0	115.0
nbr.null	0.0	0.0	0.0	0.0	0.0	0.0
nbr.na	0.0	0.0	0.0	0.0	0.0	0.0
min	108.0	26.8	10.3	13.1	5.0	3045.9
max	1584.0	121.5	62.7	37.6	22.1	7358.6
range	1476.0	94.7	52.4	24.5	17.0	4312.6
sum	90443.0	7732.1	3810.3	2736.8	1341.0	556985.8
median	453.0	65.7	33.4	23.2	11.5	4842.5
mean	786.5	67.2	33.1	23.8	11.7	4843.4
SE.mean	47.9	1.8	1.2	0.5	0.4	83.8
CI.mean.0.95	94.8	3.5	2.3	1.0	0.8	166.0
var	263608.8	364.2	159.0	30.1	18.3	807250.4
std.dev	513.4	19.1	12.6	5.5	4.3	898.5
coef.var	0.7	0.3	0.4	0.2	0.4	0.2

Subsection 3. List of mapping for recategorisation of foods in chapter 7.

('non-uncovered_iuna_foodcode_mapping.csv')

NPNS food codes re-categorised to assess key free sugar sources.

code	description	category
1	Rice & pasta, flours, grains & starch	Bread & Cereals
2	Savouries	Other
3	White sliced bread & rolls	Bread & Cereals
4	Wholemeal & brown bread & rolls	Bread & Cereals
5	Other breads	Bread & Cereals
6	RTEBC	RTEBC
7	Other breakfast cerals	Bread & Cereals
8	Biscuits including crackers	Cakes & Biscuits
9	Cakes, pastries & buns	Cakes & Biscuits
10	Whole milk	Dairy
11	Low fat, skimmed & fortified milks	Dairy
12	Other milks & milk based beverages	Dairy
13	Creams	Dairy
14	Cheeses	Dairy
15	Yogurts	Dairy
16	Ice creams	Dairy
17	Desserts	Desserts & Puddings
18	Rice puddings & custards	Desserts & Puddings
19	Eggs & egg dishes	Other
20	Butter	Other
21	Low fat spreads	Other
22	Other spreading fats	Other
23	Oils (not including those used in recipes)	Other
24	Hard cooking fats	Other
25	Potatoes	Fruit & vegetables
26	Processed & homemade potato products	Fruit & vegetables
27	Chipped, fried & roasted potatoes	Fruit & vegetables

28	Vegetable & pulse dishes	Fruit & vegetables
29	Peas, beans & lentils	Fruit & vegetables
30	Green vegetables	Fruit & vegetables
31	Carrots	Fruit & vegetables
32	Salad Vegetables	Fruit & vegetables
33	Other Vegetables	Fruit & vegetables
34	Tinned or jarred vegetables	Fruit & vegetables
35	Fruit juices	Fruit juices & smoothies
36	Bananas	Fruit & vegetables
37	Other fruit	Fruit & vegetables
38	Citrus fruits	Fruit & vegetables
39	Tinned fruit	Fruit & vegetables
40	Nuts & seeds, herbs & spices	Other
41	Fish & fish products	Other
42	Fish dishes	Other
43	Bacon & ham	Other
44	Beef & veal	Other
45	Lamb	Other
46	Pork	Other
47	Chciken, turkey & game	Other
48	Offal and offal dishes	Other
49	Beef & veal dishes	Other
50	Lamb, pork & bacon dishes	Other
51	Poultry & game dishes	Other
52	Burgers (beef & pork)	Other
53	Sausages	Other
54	Meat pies & pastries	Other
55	Meat products	Other
56	Alcoholic beverages	Other
57	Sugars, syrups, preserves & sweeteners	Sugar & syrups
58	Chocolate confectionery	Chocolate confectionery
59	Non-chocolate confectionery	Non-chocolate confectionery
60	Savoury snacks	Other
61	Soups, sauces & miscellaneous foods	Other
62	Supplements	Other
63	Teas	Other
64	Coffees	Other
65	Other beverages	Other
66	Carbonated beverages	Soft Drinks (non-diet)
67	Diet carbonated beverages	Soft Drinks (Diet)

68	Squashes, cordials & fruit juice drinks	Soft Drinks (non-diet)
69	Infant cereals	Bread & Cereals
70	Infant biscuits/rusks	Bread & Cereals
71	Infant milks	Dairy
72	Fromage frais	Dairy
73	Infant desserts (excl. pureed fruit)	Desserts & Puddings
74	Infant meals, vegetable	Fruit & vegetables
75	Fruit purees & smoothies (incl veg/fruit combinations)	Fruit juices & smoothies
76	Infant meals, fish	Other
77	Infant meals, meat	Other

Appendix B

dentistry journal

MDPI

Article

Weight Status and Dental Problems in Early Childhood: Classification Tree Analysis of a National Cohort

Michael Crowe [1,*], Michael O' Sullivan [1], Oscar Cassetti [1] and Aifric O' Sullivan [2]

[1] Division of Restorative Dentistry & Periodontology, Dublin Dental University Hospital, Trinity College Dublin, Dublin, Dublin 2, Ireland; michael.osullivan@dental.tcd.ie (M.O.S.); oscar.getstring@gmail.com (O.C.)

[2] UCD Institute of Food and Health, 2.05 Science Centre, South, UCD, Belfield, Dublin, Dublin 4, Ireland; aifric.osullivan@ucd.ie

* Correspondence: michael.crowe@dental.tcd.ie; Tel.: +353-1-612-7312

Received: 22 July 2017; Accepted: 29 August 2017; Published: 31 August 2017

Abstract: A poor quality diet may be a common risk factor for both obesity and dental problems such as caries. The aim of this paper is to use classification tree analysis (CTA) to identify predictors of dental problems in a nationally representative cohort of Irish pre-school children. CTA was used to classify variables and describe interactions between multiple variables including socio-demographics, dietary intake, health-related behaviour, body mass index (BMI) and a dental problem. Data were derived from the second (2010/2011) wave of the 'Growing Up in Ireland' study (GUI) infant cohort at 3 years, n = 9793. The prevalence of dental problems was 5.0% (n = 493). The CTA model showed a sensitivity of 67% and specificity of 58.5% and overall correctly classified 59% of children. Ethnicity was the most significant predictor of dental problems followed by longstanding illness or disability, mother's BMI and household income. The highest prevalence of dental problems was among children who were obese or underweight with a longstanding illness and an overweight mother. Frequency of intake of some foods showed interactions with the target variable. Results from this research highlight the interconnectedness of weight status, dental problems and general health and reinforce the importance of adopting a common risk factor approach when dealing with prevention of these diseases.

Keywords: body mass index; diet; dental problem; classification tree

1. Introduction

Early childhood caries (ECC) is a disease defined by the presence of one or more decayed, missing (due to caries), or filled tooth surfaces in any primary tooth in a child < 71 months old [1]. ECC is the most prevalent dental problem in pre-schoolers [2], one of the most common causes of hospital admission and the most common cause of dental extractions under general anaesthesia [1,2]. Obesity, defined as an excess of body fat [3], is another growing concern among preschool children. Body mass index (BMI) is frequently used to classify adults as overweight or obese; however, classifying overweight and obesity in children is complicated by age and gender specific differences [3,4]. For this reason, the International Obesity Task Force (IOTF) defines childhood weight status based on BMI centile curves that correspond to adult criteria from 2 to 18 years for males and females [5]. In Europe, 12%–15% of preschool children are classified as overweight or obese based on IOTF criteria [6]. Concerns around EEC and childhood obesity are heightened by the fact that both are strong predictors of these respective conditions throughout the life-course [7,8].

The preschool age is a particularly important period to minimise the risks for dental caries and obesity [9] and the primary caregiver (PCG) plays a key role in facilitating prevention through

feeding patterns and other behaviours [1,8,10]. Obesity and dental caries share some common risk factors including food choice, dietary intake patterns, diet quality and socioeconomic factors such as PCG education and household income [11–13]. Given the associations that exist between oral health and general health interest is growing in using a common risk factor approach to investigate the multidimensional causes of dental and weight-status problems, particularly in preschool children [14–17]. Although some studies have shown a positive relationship between BMI and dental caries, others suggest that they are weakly correlated, inverse or even U-shaped and that different predictors may be associated with dental caries at both high and low BMI levels [11,12,14,17]. Indeed, very few studies report the oral health status of underweight children and often group underweight and normal weight without considering differences in risk [14,18].

Data-driven methods are being increasingly proposed to empirically derive dietary patterns associated with chronic disease [19]. Methods that aim to uncover the relationship between independent variables and a dependent variable are described as supervised learning. The discovered relationship is typically presented as a classification or regression model [20]. Thus, in Classification and Regression Tree Analysis (CART) when the target (dependent) variable is continuous a regression analysis is performed and when the target variable is categorical a classification tree analysis (CTA) is carried out. Data mining techniques are invaluable when analysing multidimensional data from large-scale survey microdata files as they provide a means to identify novel diet-disease relationships and can help establish inter-relationships between causal factors [20].

With a few exceptions, most national dental surveys tend to focus on children aged 5 and older. While nationally representative studies of obesity prevalence in older Irish children are well documented [21] there are few, apart from a National Preschool Nutrition Survey [22] that relate to pre-schoolers. The research in this secondary analysis proposed to use a flexible analytical approach (CTA) to explore the multilevel relations between weight status and dental problems in a large, nationally representative cohort of 3-year old children from the 'Growing Up in Ireland' study (GUI).

2. Methods

2.1. Data Collection and Participants

The aim of the GUI infant survey is to determine the individual, family and wider social and environmental factors that affect the development of children. GUI is a nationally representative longitudinal study that collected data from infants at 9 months (Wave-1) and followed up when children were 3-years old (Wave-2) providing the data file used in this study with 9793 cases. Between December 2007 and June 2008 GUI selected a random sample on a systematic basis, pre-stratified by marital status, county of residence, nationality and number of children from the National Child Benefits Register which is a universal welfare entitlement in the Republic of Ireland [23]. The sampling fraction was 0.42 with an overall response of 64.5% providing 11,134 families in Wave 1. Follow-up interviews for Wave-2 occurred between December 2010 and July 2011 and 91% of families responded while 3.8% emigrated or deceased and the remainder either refused or were not contactable. The PCG was interviewed in the family home using a questionnaire after written informed consent was obtained [23].

2.2. Anthropometric Measurements

A standard (Leicester) portable height stick was used to measure height of PCGs and children. The weight of the children was recorded using a digital scales (SECA 835, Hamburg, Germany). Data for height and weight of the PCGs were used from Wave-1 measurements and only taken at Wave-2 if they were missing or required rechecking. PCG weight was recorded using a flat mechanical scale (SECA 761, Hamburg, Germany).

BMI was calculated as weight divided by height squared (kg/m^2) and, for children, classified as overweight, obese, normal weight or thinness according to the IOTF age and gender specific

cut-offs for 3-year olds [5,24]. For simplicity, the classification of thinness (low BMI for age) is also described as underweight in this paper although the latter strictly means low weight for age in children. Overweight and obesity for children was also classified using the UK adaptation of percentile cut-offs from the WHO Multicentre Growth Reference Study (MGRS) with overweight criteria defined as a BMI between the 91st and 98th percentile while obesity was defined as a BMI on or greater than the 98th percentile [25]. PCG BMI was categorised into underweight (BMI < 18.5), normal (BMI 18.5–24.9), overweight (BMI 25–29.9) and obese (BMI > 30).

2.3. CTA Target Variable

The dichotomous target variable was a PCG reported dental problem. The question asked was: Has <child> been to visit the dentist because of a problem with his/her teeth?

2.4. CTA Predictor Variables

Attributes (independent variables) that were relevant to the target variable (dependent variable) were selected for inclusion in the model based on findings from previous research. The demographic and socioeconomic variables selected were child gender, PCG age and gender, ethnicity, PCG education level, family social class and annual equivalised household income [1,14–16,26,27]. Ethnicity was defined as Irish, Any other White background, Black, Asian or Other. The highest education level attained by the PCG was one of thirteen categories ranging from no formal education to doctorate level which was collapsed to five groups for descriptive analysis. Family social class was measured using the Irish Central Statistics Office's classification based on occupation, categorising families into one of seven groups which was collapsed to four groups for descriptive analysis. Annual disposable household income was calculated by using an equivalence scale to "weight" each household for differences in size and composition with respect to number of adults and children [23]. Markers of health status [12,13,26,28] included PCG reported child illness, disability, allergies and injuries, as well as TV-viewing hours, tooth-brushing, soother/thumb-sucking, and breastfeeding as markers of health behaviour [9,14,26,29]. Dietary intake [9,11,26] was assessed using a modified version of the Sallis-Amherst Food Frequency Questionnaire from the Longitudinal Study of Australian Children (LSAC) [30]. PCG reported the child's frequency of consumption of 15 food categories (e.g., 'sweets', 'fizzy drinks/minerals/cordials') over the previous 24-h as once, more than once or none at all.

2.5. Data Analysis

Wave-1 GUI data were statistically re-weighted to represent the population. Wave-2 data was weighted for attrition between waves and emigration combined with the Wave-1 weight [31].

Classification tree analysis (CTA) is based on recursive partitioning whereby the algorithm repeatedly creates splits in the sample based on the most significant predictor variable. The root node contains the entire sample and each subsequent split results in child nodes with the proportion of the classes displayed together with an adjusted P value. For this analysis the following parameters were selected in either SPSS (v. 20.0: SPSS, Chicago, IL, USA) or SPSS modeller (IBM SPSS Modeler v. 14.2: Chicago, IL, USA) using the Chi-squared Automatic Interaction Detection (CHAID) algorithm [32]: maximum tree depth = 5, parent node = 100, child node = 50 and bonferroni-adjusted chi-square statistic, significance < 0.05. A 10-fold cross-validation assessed model performance and produced an average misclassification risk. Details of the analysis methods were previously reported [33]. The degree of missing cases in the 3-year old GUI infant cohort was small except for the PCG BMI (5.2%), equivalised annual income (5.5%) and child BMI (2.6%), as previously reported [29]. The CHAID algorithm handles missing values by defining a separate category and treating them as a single category so that they are not excluded in the analysis [34]. A binary logistical regression analysis (*forward-wald*) was also conducted to compare findings with those generated by the classification tree output. A confusion matrix for a binary classifier provided estimation of selected performance metrics.

3. Results

3.1. Cohort Profile

Five percent of 3-year olds had a dental problem. As is common in investigations of health outcome the class distribution of the dataset was imbalanced. The minority class was the positive instances of having 'a dental problem' and the negative response was the majority class.

Table 1 describes the cohort characteristics, including anthropometric measurements, child health and behaviours. Almost all of the self-identified PCGs were female and the biological parent of the study child. Eighty five percent were 'Irish'. Using the IOTF cut-offs [5] the prevalence of thinness and obesity were 5.7% each with an additional ~18% of children being overweight. Using the WHO growth charts and BMI cut-offs, the prevalence of overweight was 18.5% and obesity was 12.8%.

Table 1. Weighted [a] Sample Characteristics, Growing Up in Ireland infant cohort participants 2010/11 (Child 3-years of age).

	Child		PCG	
			Mean	**SD**
Age (years)			29.6	(6.1)
Gender	***n***	%	***n***	%
Male	5024	51.3	161	1.6
Female	4769	48.7	9632	98.4
Anthropometrics	**Mean**	**SD**	**Mean**	**SD**
Weight (Kg)	15.27	(2.02)		
Height (m)	95.48	(3.92)		
Body Mass Index (Kg/m^2)				
Total	16.71	(1.61)	25.99	(5.16)
Male	16.99	(1.52)	26.98	(5.59)
Female	16.71	(1.61)	25.97	(5.15)
BMI Categories	***n***	%	***n***	%
Thinness IOTF	557	5.7	166	1.7
Normal IOTF	6685	68.3		
Normal WHO	6464	66.0	4523	46.2
Overweight IOTF	1737	17.7		
Overweight WHO	1815	18.5	2941	30.0
Obese IOTF	559	5.7		
Obese WHO	1257	12.8	1655	16.9
Missing	256	2.6	508	5.2
Child Health and Behaviours	***n***	%		
Dental Problems (in last 12 months)	493	5		
Longstanding illness or disability	1543	15.8		
Hospital admission (ever)	1569	16.1		
Tooth brushing 2 or more per day	5107	52.2		
Tooth brushing <2 per day	4685	47.8		
Thumb sucking	765	7.8		
Soother	3163	32.3		
TV viewing time (min/day)	1133	(72.0)		
TV viewing 1 hour or less per day	3569	36.4		
TV viewing 2 hours or less per day	3587	36.6		
TV viewing 2 hours or more per day	2630	26.9		

Table 1. *Cont.*

Child	PCG	
	Mean	**SD**
Socio-Demographics	***n***	%
Ethnicity		
Irish	8261	84.4
Non-Irish white	1018	10.4
Black	252	2.6
Asian	202	2.1
Other	54	0.6
Family Social Class		
Professional/Managerial	4553	46.5
Other non-manual/Skilled manual	3233	33.0
Semi-skilled/Unskilled	1061	10.8
Unclassified	947	9.7
Highest Education Level		
Lower secondary or less	1361	13.9
Upper secondary	3192	32.6
Non-degree	2080	21.2
Third level	3144	32.1
	Mean	**SD**
Equivalised Annual Income (€)	18,004	(10,997)

Data presented as mean and standard deviation (SD) or n and percentage. [a] Sample weighting factors applied to statistically adjust the data to be more representative of the population. IOTF, International Obesity Task Force; WHO, World Health Organisation.

The frequency of food items consumed are reported in Figure 1. The majority of children consumed water' (~83.0%), 'full-fat milk/cream' (~84.5%), 'full-fat cheese/yoghurt' (~85.0%), 'cooked veg' (~85.0%), 'fresh fruit' (~89%), and 'biscuits/doughnuts/cake/chocolate' (~74%) once or more than once in the previous 24-h. Of interest, a considerable proportion of 3-year olds consumed "un-healthy" foods including 'crisps' (~47%), "hot-chips" (~28%), sugar containing drinks (~30%), and sweets (~49%).

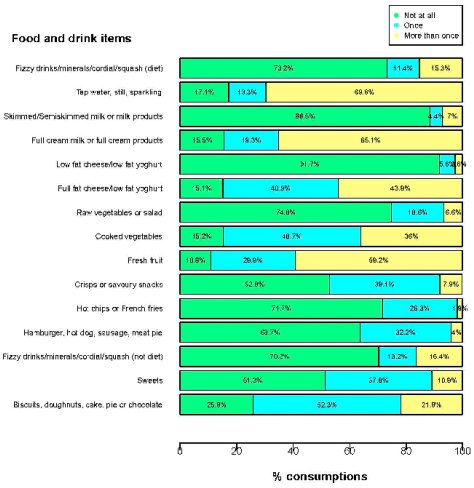


Figure 1. Food and drink items consumed in previous 24 h by the Growing Up in Ireland infant cohort at 3-years of age.

3.2. *Classification Tree Analysis*

CHAID analysis generated a CTA output as depicted in Figure 2 with 30 nodes, including 17 terminal nodes. Each node contains the number and percentage of infants in each category for the dependant variable (dental problem), the categories chosen by CHAID for the predictor variable and the cut-off points for continuous variables. PCG ethnicity was the most important predictor of the 3-year old child having a dental problem splitting the root node. Twelve predictor variables were included in the final tree (Bonferroni-adjusted $p < 0.05$). Two predictors appeared twice in the output, PCG BMI (nodes 2 and 5) and equivalised household annual income (nodes 3 and 4). A confusion matrix (Table 2) produced performance metrics for the classification tree: sensitivity 66.8%, specificity 58.5% and overall accuracy 58.9%.

The ethnic subgroups were split into 3 nodes with the highest prevalence of dental problems (8.4%) among those children from a "non-Irish white" background (Node-3). Node-1 contained almost 87% of the sample (Irish and Asian ethnicity) with a 4.7% prevalence of dental problems. The tree output from node-1 to nodes 22–24 delineated subgroups linking child BMI categories with dental problems by the following predictors: PCG from an Irish/Asian background (node-1), the presence of a longstanding illness or disability in the child (node-5) and an overweight mother (node-13). The final predictor at node-13 was BMI classification of the child which split into three terminal nodes resulting in normal, overweight/missing and obese/underweight subgroups. The highest dental problem prevalence (19%, $n = 17$) was in those children in this final subgroup who were obese or underweight (node-24). Also, the subgroup at node-1 who had a longstanding illness or disability had a reported dental problem prevalence of 7.0% while those with no illness or disability had a prevalence of 4.3% (node-4). The food variables included in the tree output were 'water' (level-3), 'low-fat cheese/yoghurt' (level-4) and 'raw vegetables/salad', 'fresh fruit' and 'hot chips'.

Logistic regression failed to generate a significant model (chi-square (6) = 9.38, $p = 0.15$).

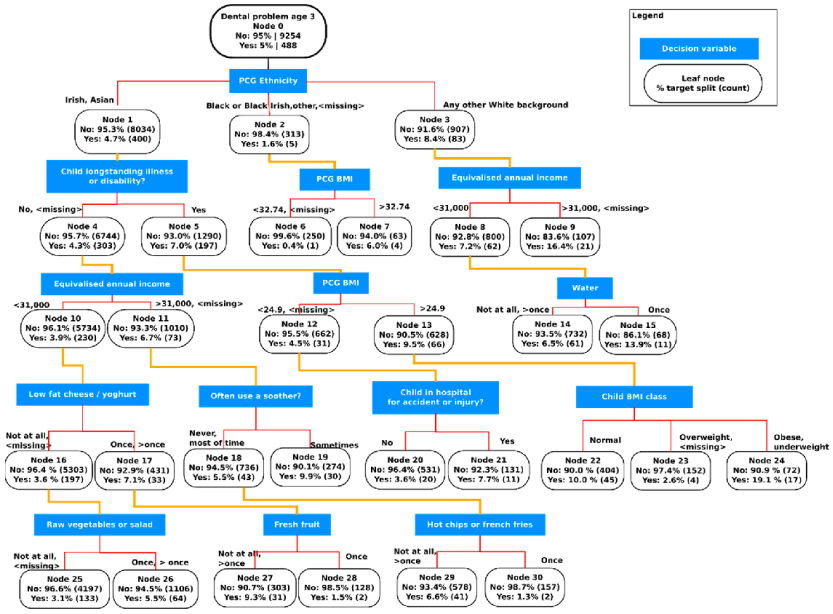


Figure 2. Prevalence of reported dental problems by the Growing Up in Ireland infant cohort at 3-years of age among classification tree subgroups, percentage (%) and number (n) in each class.

Table 2. Confusion matrix showing selected performance measures for Classification tree analysis of dental problem prevalence in the Growing Up in Ireland infant cohort at 3-years of age.

Observed		Predicted Dental Problem		Percentage Correct	Measure
		Yes	No		
Dental problem	Yes	326	162	66.8%	Sensitivity [a]
	No	3839	5415	58.5%	Specificity [b]
Overall percentage				58.9%	Accuracy [c]

[a] Sensitivity = True Positive Rate = Number of True Positives/(Number of True Positives + Number of False Negatives); [b] Specificity = True Negative Rate = Number of True Negatives/(Number of True Negatives + Number of False Positives); [c] Accuracy = (True Positives + True Negatives)/(True Positives + False Positives + False Negatives + True Negatives)

4. Discussion

This study used a CHAID classification tree as a method to classify the dataset and identify relationships between the predictor variables selected and the target variable (a PCG-reported dental problem requiring a visit to the dentist). PCG ethnicity was the most significant predictor of dental problems in the CTA model and the highest prevalence of dental problems in this study was among children who were obese or underweight with a longstanding illness and an overweight PCG.

This analysis was carried out using data from a nationally representative cohort of 3-year old children, the largest child population study ever carried out in Ireland, which includes a wide range of PCG and child health and development characteristics. CTA is a non-parametric method which handles nominal and numeric input and classification trees are ideal for representing complex interactions [20]. The output produces a visualisation of all the significant interactions with the target variable at multiple levels and can, potentially, uncover subgroups that might not be discovered using other data methods [34]. The partitioning variable at the first level of the classification tree was PCG ethnicity while at the next level the most significant predictors of dental problems were the child having a longstanding illness, PCG BMI and household income. It has been reported that trends in overweight and obesity differ among different ethnic groups, even at the early preschool age, and that this cannot be explained by variations in household income [35]. Similarly, the disparities in dental caries prevalence between different ethnic groups is not fully explained by social inequalities [27]. Surprisingly, ethnicity has been used as a variable in relatively few studies of dental caries in children while the PCG education level, although it was not a significant predictor in our CTA model, has been consistently shown to be an important risk factor for caries in children [26].

While the results of our CTA are exploratory they identify certain characteristics previously suggested as risk factors or risk indicators for both obesity and dental caries [14,26,36]. The study supports recent views that data driven-outcome dependant methods such as CTA are potentially useful for investigating dietary components or patterns most associated with a health outcome and are a valid, non-parametric, alternative to logistic regression analysis [19]. The results of this analysis must be cautiously interpreted by gauging the model performance (Table 2) and understanding the limitations of both the data structure and classification tree algorithms. An imbalanced class distribution has been characterised as "one that has many more instances of some classes than others [37]. CTA of imbalanced data sets tends to result in high predictive accuracy for the majority class and low accuracy for the minority class [34]. In most health outcome investigations, including dental problems, the correct classification of the minority class is of greater interest or value than that of the majority class. The confusion matrix (Table 2) shows the results of the actual and predicted classifications carried out by CTA. The metrics calculated include sensitivity or recall (66.8%) which is the proportion of actual positive cases correctly predicted by the model and specificity (58.5%) which is the proportion of actual negative cases correctly identified by the model. The overall accuracy (58.9%) indicated the proportion of the total number of correct predictions. Logistic regression did not perform well as a classifier and none of the same input variables were significant in the final

regression model. This may be due to inherent differences in the way CTA captures the division of the classes by partitioning the space using multiple decision boundaries whereas logistic regression uses a single linear decision boundary [34]. To be suitable as a prediction model for targeting risk it has been suggested that both sensitivity and specificity should be 80% or the sum be at least 160% [38].

While overweight and obesity dominate the focus of recent research with children, it is important to consider underweight (thinness) in early childhood as a condition related to poor health outcomes also. There is some evidence to suggest that dental caries may be associated with children who are underweight and suffer with slow growth due to pain on mastication [15,18,28]. The results (Table 1) shows a similar prevalence of underweight and obese children. A small subgroup (node-24) of children which combined obese and underweight categories had the highest prevalence of dental problems (19%) in the sample. This group were predominantly Irish with a longstanding illness and had an overweight PCG. It should also be noted that normal weight children (node-22) in this group had a dental problem prevalence of 10% approximately half that of the obese/underweight group, but double that of the overall sample. This finding of itself highlights the interconnectedness of weight status, dental problems and general health and reinforces the importance of adopting a common risk factor approach when dealing with prevention of these diseases [13,39]. However, while of interest in classifying this dataset, it is important to be cautious when interpreting these subgroups identified by CTA as hierarchical splitting means that they are mutually exclusive. Furthermore, successful targeting of high risk population subgroups for problems with both weight status and dental health would require a risk prediction model with both high sensitivity and specificity.

The prevalence of PCG-reported dental problems requiring a visit to the dentist was 5% which may be an under-estimate given that dental problems are often not treated unless symptomatic in the preschool years [28]. This age is a pivotal period for development of both obesity and dental caries as patterns of eating behaviour that predispose to later development of these conditions are established [8,10,15]. The prevalence of overweight or obesity in 3-year old children determined by IOTF cut-offs was approximately 23% which was similar to previous reports [22]. Almost 47% of PCG's were overweight or obese and it is well established that parental overweight and obesity increases the risk of a child becoming overweight [39]. There are limitations in using BMI as an indirect measure of "fatness" particularly with respect to children [4] and it is important to note that there is no reference population in Ireland for grading BMI. In the CTA, we used the IOTF classification as the more conservative estimate of obese children with a higher cut-off threshold. The FFQ adopted for the GUI survey was a modified dietary screening recall and provided an indication of types and frequency of foods consumed. While PCG-reported measures of foods consumed on a single occasion may be useful in differentiating patterns of food intake it does not provide a good estimate of usual daily consumption and cannot accurately capture total energy or total nutrient intake [40]. Preschool children with unhealthy eating habits have an increased likelihood of experiencing dental caries [10]. While obesity and dental caries are both diet-mediated diseases it is clear that sugars are required in the diet for dental caries to occur [41] whereas a high consumption of energy dense foods including sugars and saturated fats are linked with obesity [16,39]. Fundamentally, obesity occurs due to an energy imbalance between calories consumed versus those expended over a period of time [11,39]. Approximately 74% of the children in GUI consumed biscuits, doughnuts, cake, pie or chocolate at least once or more than once in a 24-h period (Figure 1). Almost 49% of children ate sweets and 30% drank non-diet fizzy drinks, minerals, cordials or squash at least once or more than once. Dietary interventions aimed at reducing the intake of these unhealthy food groups may help impact on both obesity and dental caries but further investigation of these factors in longitudinal studies is still required, especially given the temporal and cumulative aspects of both conditions.

The results of this data analysis may help raise awareness among clinicians and nutrition researchers of the interrelations between weight status and dental problems, even before the primary dentition is complete. Classification tree analysis visually demonstrates how factors, such as the BMI of the primary caregiver (PCG), can interact at multiple levels and affect different subgroups of the

child population. Future intervention strategies for oral health should involve consideration of the weight of both the young child and PCG at both the patient and population level. This approach has been advocated recently as both obesity and dental caries may be more likely to occur in the same populations [17]. Given the increasing public health implications of these conditions adopting a more interdisciplinary approach to shared risk factors may assist in the reduction of both problems.

A limitation of CTA is that the hierarchical nature of recursive partitioning has an inherent instability to small changes in the learning data [34]. Overfitting of the model is a known problem and this can be guarded against by using cross-validation which splits the dataset into a training portion and test portion before estimating an average misclassification risk [20]. While dental caries is the most common dental problem at this age the survey did not include a report of the outcome of the dental visit. However, it has been shown that dental disease and treatment need of young children are associated with parents' perceptions of their children's oral health status [42]. Also, it is important to note that this study was based on a cross sectional rather than longitudinal analysis of the infant cohort from the GUI survey. The data is reliant on PCG reporting and this is subject to recall bias and social desirability, particularly in relation to reporting of food and drink perceived as 'healthy' or 'unhealthy'. Further research is focussed on maximising the quality of food intake data by augmenting the GUI survey data with more reliable dietary intake values from a national nutritional database [43] and using the longitudinal aspect of the next wave of the GUI infant cohort. Inclusion of more detailed oral health measures should also be considered.

5. Conclusions

The highest prevalence of dental problems in this study was among children who were obese or underweight with a longstanding illness and an overweight PCG. Societal changes may require renewed focus on oral health policies to focus on minority groups and CTA is a novel approach for exploring large survey data and health-related outcomes. The common risk factor approach may be a pragmatic means of developing shared modifiable strategies for prevention of both dental and weight problems.

Acknowledgments: Data have been collected under the Statistics Act, 1993, of the Central Statistics Office. The GUI survey has been designed and carried out by the ESRI-TCD Growing up in Ireland team.

Author Contributions: All authors contributed equally to this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gussy, M.G.; Waters, E.G.; Walsh, O.; Kilpatrick, N.M. Early childhood caries: Current evidence for aetiology and prevention. *J. Paediatr. Child Health* **2006**, *42*, 37–43. [CrossRef] [PubMed]
2. Public Health England. *National Dental Epidemiology Programme for England: Oral Health Survey of Five-Year-Old Children 2012*; A Report on the Prevalence and Severity of Dental Decay; Public Health England: London, UK, 2013. Available online: http://www.nwph.net/dentalhealth/survey-results5.aspx?id=1 (accessed on 20 February 2017).
3. Flegal, K.M.; Ogden, C.L. Childhood obesity: Are we all speaking the same language? *Adv. Nutr.* **2011**, *2*, 159S–166S. [CrossRef] [PubMed]
4. Rolland-Cachera, M.F. Childhood obesity: Current definitions and recommendations for their use. *Int. J. Pediatr. Obes.* **2011**, *6*, 325–331. [CrossRef] [PubMed]
5. Cole, T.J.; Bellizzi, M.C.; Flegal, K.M.; Dietz, W.H. Establishing a standard definition for child overweight and obesity worldwide: International survey. *BMJ* **2000**, *320*, 1240. [CrossRef] [PubMed]
6. Ahrens, W.; Pigeot, I.; Pohlabeln, H.; de Henauw, S.; Lissner, L.; Molnár, D.; Moreno, L.; Tornaritis, M.; Veidebaum, T.; Siani, A. Prevalence of overweight and obesity in European children below the age of 10. *Int. J. Obes.* **2014**, *38*, S99–S107. [CrossRef] [PubMed]
7. Dye, B.A.; Hsu, K.L.; Afful, J. Prevalence and measurement of dental caries in young children. *Pediatr. Dent.* **2015**, *37*, 200–216. [PubMed]

8. Wake, M.; Hardy, P.; Sawyer, M.G.; Carlin, J.B. Comorbidities of overweight/obesity in australian preschoolers: A cross-sectional population study. *Arch. Dis. Child.* **2008**, *93*, 502–507. [CrossRef] [PubMed]
9. Chaffee, B.W.; Feldens, C.A.; Rodrigues, P.H.; Vitolo, M.R. Feeding practices in infancy associated with caries incidence in early childhood. *Commun. Dent. Oral Epidemiol.* **2015**, *43*, 338–348. [CrossRef] [PubMed]
10. Dye, B.A.; Shenkin, J.D.; Ogden, C.L.; Marshall, T.A.; Levy, S.M.; Kanellis, M.J. The relationship between healthful eating practices and dental caries in children aged 2–5 years in the united states, 1988–1994. *J. Am. Dent. Assoc.* **2004**, *135*, 55–66. [CrossRef] [PubMed]
11. Marshall, T.A.; Eichenberger-Gilmore, J.M.; Broffitt, B.A.; Warren, J.J.; Levy, S.M. Dental caries and childhood obesity: Roles of diet and socioeconomic status. *Commun. Dent. Oral Epidemiol.* **2007**, *35*, 449–458. [CrossRef] [PubMed]
12. Kantovitz, K.R.; Pascon, F.M.; Rontani, R.M.P.; Gaviao, M.B.D. Obesity and dental caries—A systematic review. *Oral Health Prev. Dent.* **2006**, *4*, 137–144. [PubMed]
13. Sheiham, A.; Watt, R.G. The common risk factor approach: A rational basis for promoting oral health. *Commun. Dent. Oral Epidemiol.* **2000**, *28*, 399–406. [CrossRef]
14. Hooley, M.; Skouteris, H.; Boganin, C.; Satur, J.; Kilpatrick, N. Body mass index and dental caries in children and adolescents: A systematic review of literature published 2004 to 2011. *Syst. Rev.* **2012**, *1*, 57. [CrossRef] [PubMed]
15. Hooley, M.; Skouteris, H.; Millar, L. The relationship between childhood weight, dental caries and eating practices in children aged 4–8 years in australia, 2004–2008. *Pediatr. Obes.* **2012**, *7*, 461–470. [CrossRef] [PubMed]
16. Hayden, C.; Bowler, J.O.; Chambers, S.; Freeman, R.; Humphris, G.; Richards, D.; Cecil, J.E. Obesity and dental caries in children: A systematic review and meta-analysis. *Commun. Dent. Oral Epidemiol.* **2012**, *41*, 289–308. [CrossRef] [PubMed]
17. Public Health England. *The Relationship between Dental Caries and Obesity in Children: An Evidence Summary*; Public Health England: London, UK, 2015. Available online: https://www.gov.uk/government/publications/dental-caries-and-obesity-their-relationship-in-children (accessed on 5 March 2017).
18. Tramini, P.; Molinari, N.; Tentscher, M.; Demattei, C.; Schulte, A.G. Association between caries experience and body mass index in 12-year-old french children. *Caries Res.* **2009**, *43*, 468–473. [CrossRef] [PubMed]
19. Krebs-Smith, S.M.; Subar, A.F.; Reedy, J. Examining dietary patterns in relation to chronic disease: Matching measures and methods to questions of interest. *Circulation* **2015**, *132*, 790–793. [CrossRef] [PubMed]
20. Yoo, I.; Alafaireet, P.; Marinov, M.; Pena-Hernandez, K.; Gopidi, R.; Chang, J.F.; Hua, L. Data mining in healthcare and biomedicine: A survey of the literature. *J. Med. Syst.* **2012**, *36*, 2431–2448. [CrossRef] [PubMed]
21. Keane, E.; Kearney, P.M.; Perry, I.J.; Kelleher, C.C.; Harrington, J.M. Trends and prevalence of overweight and obesity in primary school aged children in the republic of ireland from 2002–2012: A systematic review. *BMC Public Health* **2014**, *14*, 974. [CrossRef] [PubMed]
22. Walton, J. *National Pre-School Nutrition Survey*; Summary Report on: Food and Nutrient Intakes, Physical Measurements and Barriers to Healthy Eating; Irish Universities Nutrition Alliance: Dublin, Ireland, 2012. Available online: http://www.iuna.net/ (accessed on 11 March 2017).
23. Murray, A.; Quail, A.; McCrory, C.; Williams, J. *A Summary Guide to Wave 2 of the Infant Cohort (at 3 Years) of Growing up in Ireland*; The Economic and Social Research Institute: Dublin, Ireland, 2013.
24. Cole, T.J.; Flegal, K.M.; Nicholls, D.; Jackson, A.A. Body mass index cut offs to define thinness in children and adolescents: International survey. *BMJ* **2007**, *335*, 194. [CrossRef] [PubMed]
25. Cole, T.J.; Wright, C.M.; Williams, A.F.; Group, R.G.C.E. Designing the new uk-who growth charts to enhance assessment of growth around birth. *Arch. Dis. Child. Fetal Neonatal Ed.* **2012**, *97*, F219–F222. [CrossRef] [PubMed]
26. Harris, R.; Nicoll, A.D.; Adair, P.M.; Pine, C.M. Risk factors for dental caries in young children: A systematic review of the literature. *Commun. Dent. Health* **2004**, *21*, 71–85.
27. Van der Tas, J.T.; Kragt, L.; Veerkamp, J.J.; Jaddoe, V.W.; Moll, H.A.; Ongkosuwito, E.M.; Elfrink, M.E.; Wolvius, E.B. Ethnic disparities in dental caries among six-year-old children in the netherlands. *Caries Res.* **2016**, *50*, 489–497. [CrossRef] [PubMed]
28. Sheiham, A. Dental caries affects body weight, growth and quality of life in pre-school children. *Br. Dent. J.* **2006**, *201*, 625–626. [CrossRef] [PubMed]

29. Layte, R.; Bennett, A.; McCrory, C.; Kearney, J. Social class variation in the predictors of rapid growth in infancy and obesity at age 3 years. *Int. J. Obes.* **2014**, *38*, 82–90. [CrossRef] [PubMed]
30. Sallis, J.F.; Taylor, W.C.; Dowda, M.; Freedson, P.S.; Pate, R.R. Correlates of vigorous physical activity for children in grades 1 through 12: Comparing parent-reported and objectively measured physical activity. *Pediatr. Exer. Sci.* **2002**, *14*, 30. [CrossRef]
31. Quail, A.; Williams, J.; McCrory, C.; Murray, A.; Thornton, M. *Sample Design and Response in Wave 1 of the Infant Cohort (at 9 months) of Growing up in Ireland*; Department of Health and Children: Dublin, Ireland, 2011.
32. Kass, G.V. An exploratory technique for investigating large quantities of categorical data. *Appl. Stat.* **1980**, *29*, 119–127. [CrossRef]
33. Crowe, M.; O'Sullivan, A.; McGrath, C.; Cassetti, O.; Swords, L.; O'Sullivan, M. Early childhood dental problems classification tree analyses of 2 waves of an infant cohort study. *JDR Clin. Trans. Res.* **2016**, *1*, 275–284. [CrossRef]
34. Maimon, O.; Rokach, L. Classification trees. In *Data Mining and Knowledge Discovery Handbook*; Maimon, O., Rokach, L., Eds.; Springer: New York, NY, USA, 2005; Volume 2, pp. 149–174.
35. Karlsen, S.; Morris, S.; Kinra, S.; Vallejo-Torres, L.; Viner, R.M. Ethnic variations in overweight and obesity among children over time: Findings from analyses of the health surveys for england 1998–2009. *Pediatr. Obes.* **2014**, *9*, 186–196. [CrossRef] [PubMed]
36. De Onis, M.; Lobstein, T. Defining obesity risk status in the general childhood population: Which cut-offs should we use? *Int. J. Pediatr. Obes.* **2010**, *5*, 458–460. [CrossRef] [PubMed]
37. Sun, Y.; Kamel, M.S.; Wong, A.K.C.; Wang, Y. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognit.* **2007**, *40*, 3358–3378. [CrossRef]
38. Hausen, H. Caries prediction–state of the art. *Commun. Dent. Oral Epidemiol.* **1997**, *25*, 87–96. [CrossRef]
39. Lobstein, T.; Jackson-Leach, R.; Moodie, M.L.; Hall, K.D.; Gortmaker, S.L.; Swinburn, B.A.; James, W.P.T.; Wang, Y.; McPherson, K. Child and adolescent obesity: Part of a bigger picture. *Lancet* **2015**, *385*, 2510–2520. [CrossRef]
40. Magarey, A.; Watson, J.; Golley, R.K.; Burrows, T.; Sutherland, R.; McNaughton, S.A.; Denney-Wilson, E.; Campbell, K.; Collins, C. Assessing dietary intake in children and adolescents: Considerations and recommendations for obesity research. *Int. J. Pediatr. Obes.* **2011**, *6*, 2–11. [CrossRef] [PubMed]
41. Peres, M.A.; Sheiham, A.; Liu, P.; Demarco, F.F.; Silva, A.E.; Assuncao, M.C.; Menezes, A.M.; Barros, F.C.; Peres, K.G. Sugar consumption and changes in dental caries from childhood to adolescence. *J. Dent. Res.* **2016**, *95*, 388–394. [CrossRef] [PubMed]
42. Talekar, B.S.; Rozier, R.G.; Slade, G.D.; Ennett, S.T. Parental perceptions of their preschool-aged children's oral health. *J. Am. Dent. Assoc.* **2005**, *136*, 364–372. [CrossRef] [PubMed]
43. Crowe, M.; O'Sullivan, M.; Cassetti, O.; McGrath, C.; O'Sullivan, A. Data mapping to augment dietary intake values from a nutritional database to a national cohort survey: Protocols to improve quality of reported food intake. *Proc. Nutr. Soc.* **2016**, *75*. [CrossRef]

ORIGINAL REPORT: EPIDEMIOLOGIC RESEARCH

Early Childhood Dental Problems: Classification Tree Analyses of 2 Waves of an Infant Cohort Study

M. Crowe[1], A. O'Sullivan[2], C. McGrath[3], O. Cassetti[1], L. Swords[4], and M. O'Sullivan[1]

Abstract: *Investigations into the wider bioecological understanding of dental problems in early childhood are limited in national surveys. Classification tree analysis (CTA) was used to explore multilevel interactions among key aspects of child and primary caregiver (PCG) psychosocial and physical health affecting dental problems in preschool children. Data were derived from the Growing Up in Ireland study, a nationally representative sample of 9-mo-olds (N = 11,134) in 2007/2008 followed up at age 3 y (N = 9,793) in 2010/2011. Analysis included PCG reports of children's dental problems, general health, temperament, emotional and behavioral difficulties, and their own general health, stress and depression, relationship, and sociodemographic variables. Misclassification costs were specified for the model by applying a higher penalty for misclassifying those with a dental problem (minority class). Logistic regression analyses were carried out for comparison. Dental problems were reported among 302 infants (2.7%) at 9 mo of age and 493 children (5.0%) at 3 y. CTA identified infant temperament (Infant Characteristics Questionnaire unpredictable) as the primary predictor of dental problems at 9 mo and child global health at 3 y of age. First-level predictors were PCG depression score and use of a soother at 9 mo and PCG ethnicity and unscheduled hospital visits at 3 y of age. Regression analyses results supported the most important predictors at 9 mo and 3 y of age. The CTA model for 9-mo-old infants had a specificity of 90.4%, sensitivity of 31.2%, and overall accuracy of 88.8% while that for 3-y-olds had a specificity of 58.5%, sensitivity of 66%, and overall accuracy of 59%. Key aspects of infant/child and PCG health, as well as psychosocial characteristics associated with reported dental problems, should be considered in future multidisciplinary approaches to child health.*

Knowledge Transfer Statement: *The results of this data analysis should help raise awareness among clinicians of how primary caregiver and child psychosocial and general health factors are associated with early childhood dental problems, even before the primary dentition is complete. Classification tree analysis visually demonstrates how factors such as infant temperament (9 mo) and child global health (3 y) can interact at multiple levels and affect different subgroups of the child population. Future intervention strategies for oral health should involve consideration of the psychological and general health characteristics of the young child and PCG at both the patient and population levels. This knowledge could assist decision makers adopt an integrated multidisciplinary approach in formulating a coherent oral health policy for preschool children.*

Keywords: oral health, psychosocial factors, temperament, global health, preschool child, primary caregiver

DOI: 10.1177/2380084416651834. [1]Division of Restorative Dentistry & Periodontology, Dublin Dental University Hospital, Trinity College Dublin, Dublin, Ireland; [2]UCD Institute of Food and Health, 2.05 Science Centre, South, UCD, Belfield, Dublin, Ireland; [3]Faculty of Dentistry, The University of Hong Kong, Hong Kong SAR, China; and [4]School of Psychology & Children's Research Centre, Trinity College Dublin, Dublin, Ireland. Corresponding author: M. Crowe, Division of Restorative Dentistry & Periodontology, Dublin Dental University Hospital, 2 Lincoln Place, Dublin 2, Dublin, Ireland. Email: michael.crowe@dental.tcd.ie

A supplemental appendix to this article is published electronically only at http://jdrctr.sagepub.com/supplemental.

Introduction

The prevalence of oral health problems in young children has increased in recent years, following a decline in previous decades (Dye et al. 2010; Bourgeois and Llodra 2014). The bidirectional relationship between oral health problems and child health and development is complicated by a variety of sociodemographic influences (Sheiham 2006; Hooley et al. 2012). In addition, the primary caregiver (PCG) is the gatekeeper in providing and promoting general and oral health care for the developing child; therefore, PCG health and well-being are intricately linked to child health and ultimately defined by similar social determinants (Moimaz et al. 2014).

In recent years, dental research has expanded, recognizing that psychosocial, behavioral, and environmental factors significantly affect oral health outcomes (Newton and Bower 2005; Fisher-Owens et al. 2007). A number of studies have reported relationships between PCG psychological distress, child socioemotional behavior or infant temperament, and child oral health outcomes (Quinonez et al. 2001; Tang et al. 2005; Spitz et al. 2006; Menon et al. 2013; Aminabadi et al. 2014). Parental stress and depression may affect caregivers' ability to impart preventive oral health measures at vulnerable developmental stages (LaValle et al. 2000; Tang et al. 2005) and are often related to aspects of infant temperament and child socioemotional behavior (Mäntymaa et al. 2006; Spitz et al. 2006; Renzaho and Silva-Sanigorski 2013). Depressive symptoms in mothers may lead to inconsistent parenting and unhealthy feeding habits (Kim Seow 2012). Furthermore, a positive child temperament appears to be protective against early childhood caries (ECC) while temperament and poor feeding practices are both equally strongly associated with ECC (Aminabadi et al. 2014). Influenced by behavioral and social science research, conceptual models of oral health have developed to incorporate a wider framework, including psychosocial and behavioral factors that expand beyond the individual to the family and community levels (Fisher-Owens et al. 2007; Kim Seow 2012). While several studies have focused on the sociodemographic components within models, few have examined the role of psychosocial and behavioral factors in large population studies (Hooley et al. 2012).

Classification and regression trees have been used in clinical settings for risk assessment or diagnostic prediction but less so in public health research (Kuhn et al. 2014). The aim of classification tree analysis (CTA) is to create a model that predicts a target outcome (dependent variable) based on the strength of interactions between categorical or continuous input variables (independent variables) (Loh 2014). To date, most of the research has focused on early childhood dental problems, and studies on the health and psychosocial attributes of the child and PCG have concentrated on the effect of a single variable using relatively small sample sizes (LaValle et al. 2000; Hooley et al. 2012). This study uses CTA to explore a complex network of infant/child and PCG psychosocial and physical health variables and identify key parameters related to early childhood dental problems in a large nationally representative sample of Irish children in infancy and again in early childhood.

Methods

This study was reported according to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines.

Participants

Growing Up in Ireland (GUI) is a nationally representative longitudinal cohort study of infants in the Republic of Ireland (ROI) (http://www.growingup.ie). The sample was randomly and systematically selected from the National Child Benefits Register prestratified by marital status, county of residence, nationality, and number of children. Families were first invited to participate when infants were 9 mo old (wave 1) and subsequently when they were 3 y old (wave 2).

Data Collection

Interviews were carried out by trained fieldworkers administering a computer-based questionnaire with the PCG in the family home after informed consent was obtained in writing (Williams et al. 2013). The project received ethical approval from a Research Ethics Committee convened by the Department of Health and Children.

Dependent Variables

The dependent variable in this analysis was a PCG-reported dental problem. At 9 mo, information on reported experience of dental problems for which the infant was taken to see a health care worker was recorded as a dichotomous variable, with a positive response indicating a dental problem. In wave 2 of the study, when the child was 3 y old, the PCG was asked, "Has the child been to visit the dentist because of a problem with his/her teeth?" Again, this was recorded as a dichotomous variable, with a positive response indicating a dental problem.

Independent Variables

Independent variables were chosen based on relevance to child dental health.

Sociodemographic variables

Socioeconomic and demographic variables selected were ethnicity and highest education level of the PCG, family social class, and equivalized household annual income (Williams et al. 2013). The sex (male/female) of the child and age and sex of the PCG was recorded.

Variables related to health

At 9 mo of age, health was assessed by PCG global ratings of infant general health and whether the infant was "admitted to a hospital ward because

of an illness or health problem." The global health rating of the infant was dichotomized as "very healthy" or "not very healthy" in a similar fashion to previous studies with preschool children (Wake et al. 2008). At 3 y of age, health was assessed by PCG global ratings of child health and whether the child "ever had an accident or injury requiring hospital treatment or admission." At both time points, PCG (global) health was assessed by self-rating on a 5-point Likert scale.

Psychosocial variables/behavioral habits

At 9 mo, infant temperament was assessed by the 24-item Infant Characteristics Questionnaire (ICQ; Bates et al. 1979) covering 4 domains: fussy-difficult, unadaptable, dull (or subdued), and unpredictable. Other questions relating to behavioral habits were whether the infant used a soother/dummy in the past week and whether the PCG ever woke the baby at night for a feed. PCG stress levels were indicated by the 18-item Parental Stress Scale (Berry and Jones 1995), and PCG depression was assessed by using the 8-item short form measure of the Centre for Epidemiological Studies Depression Scale (CES-D) (Radloff 1977). The quality of attachment (QoA) to the infant was measured using the QoA subscale from the Maternal Postnatal Attachment Scale (Condon and Corkindale 1998).

At 3 y, child behavior and emotional development were assessed using the 25-item Strengths & Difficulties Questionnaire (SDQ; Goodman 1997). Child temperament was measured using a modified version from the Longitudinal Study of Australian Children (LSAC) of the Short Temperament Scale for Toddlers (Prior et al. 2000). Questions relating to child oral behavioral habits included frequency of tooth brushing and how often the child sucked a soother or finger/thumb. The child-PCG relationship was assessed by the Child-Parent Relationship Scale (Pianta CPR-S; Berry and Jones 1995). Again, when the child was 3 y of age, PCG depression was assessed using the 8-item short form measure of the CES-D Scale, yielding a total depression score. PCGs were categorized as "depressed" or "not depressed." PCGs were also asked whether they had been "treated for depression, anxiety, or nerves."

Data Analysis

The data from the GUI infant cohort study were obtained from the Irish Social Science Data Archive (ISSDA, University College Dublin). Prior to analysis, wave 1 data were reweighted based on population statistics from the Central Statistics Office (CSO) and the National Child Benefit Register for 2008 to ensure that sample was representative of the total population. Data from wave 2 were adjusted for the children who were resident in Ireland at wave 1 but not at wave 2 and for differential response and attrition between interviews (Williams et al. 2013).

Classification trees were generated with PCG-reported experience of a dental problem at 9 mo or at 3 y of age as the target variable for each output using SPSS statistics (version 20.0: SPSS, Inc.) and the Chi-squared Automatic Interaction Detection (CHAID) algorithm (Kass 1980). Comprehensive details of our analysis technique are outlined in Appendix 1, and a recent literature review provides a detailed history of classification tree techniques (Loh 2014). Following on, a series of binary logistical regression analyses (*forward-wald*) was conducted to compare findings with those generated by CTA.

Results

Profile of the Sample

At 9 mo of age, 2.7% (n = 302) of infants had a PCG-reported dental problem for which they had sought care from a health care professional. When children were 3 y old, 5.0% (n = 493) had a dental problem for which they had sought care from a dentist (Table 1). The sociodemographic profile of the study populations was similar at both time points with the exception of the variable reflecting family income (Table 1). The mean equivalized annual household income was lower at wave 2 compared with wave 1.

Classification Tree at 9 mo of Age

The results from CTA for infants at 9 mo are shown in Figure 1, which includes the independent (predictor) variable used for each split. The tree model, without boosting or undersampling, had a sensitivity of 31.2%, a specificity of 90.4%, and an overall accuracy of 88.8%. Boosting and resampling did not greatly improve the model performance for either data set. The parameters used for growing the tree with the CHAID algorithm partitioned the data into 5 levels with 33 nodes, of which 20 were terminal nodes. Each node contains the node number, the number and percentage of infants in each category for the dependent variable (dental problem), the adjusted P value and chi-square statistic, the categories chosen by CHAID for the predictor variable, and the cutoff points for continuous variables. The 6 independent variables that reached significance in the model included 3 infant temperament subscales (ICQ fussy-difficult, dull, and unpredictable subscales), the PCG depression score, infant use of a soother/dummy, and child global health. The first level of the tree was split according to the infant temperament score (ICQ unpredictable subscale), which splits the tree root (parent node) into 4 branches (child nodes). A detailed description of the tree output is contained in Appendix 2.

In regression analyses, 3 factors were significantly associated with having a reported dental problem at 9 mo: both dull (B = 0.03, SE = 0.01, P = 0.007) and fussy-difficult (B = –0.09, SE = = 0.03, P = 0.001) temperament (ICQ subscale scores), use of a soother (B = –0.28, SE = 0.134, P = 0.04), and PCG depression score (B = 0.03, SE = 0.01, P = 0.025). Having a dull temperament or using a soother reduced the likelihood of a dental problem, while having a difficult temperament and an increased PCG depression score increased the likelihood of the infant having a dental problem.

Table 1.
Demographics and Profile of Growing Up in Ireland Infant Cohort Study Participants.

Characteristic	Total at 9 mo of Age	Total at 3 y of Age
Survey date	September 8/April 9	December 10/June 11
Sample size, *n*	11,134	9,793
Child sex, *n* (%)		
Boy	5,715 (51.3)	5,024 (51.3)
Girl	5,419 (48.7)	4,769 (48.7)
Dental problem, *n* (%)	302 (2.7)	493 (5.0)
Hospital admission (ever),[a] *n* (%)		
Yes	1,453 (13.1)	1,569 (16.1)
No	9,674 (86.9)	8,201 (83.9)
Global health child, *n* (%)		
Very healthy	9,197 (82.6)	7,312 (74.7)
Not very healthy	1,895 (17.0)	2,476 (25.3)
Infant Characteristics Questionnaire, mean (SD)		
Fussy-difficult	14.83 (5.00)	
Unpredictable	6.15 (2.66)	
Unadaptable	9.01 (3.83)	
Dull	5.84 (2.46)	
Strength and Difficulties Questionnaire, mean (SD)		7.98 (4.63)
Child-Parent Relationship (CPR-Pianta), mean (SD)		
Positive		33.77 (2.00)
Conflict		15.64 (5.42)
Short Temperament Scale (Longitudinal Study of Australian Children), mean (SD)		
Persistence		4.71 (0.82)
Sociability		4.12 (1.13)
Reactivity		2.88 (1.07)
Quality of attachment score, mean (SD)	42.50 (2.59)	
Parental stress score, stressors subscale, mean (SD)	14.64 (4.19)	12.35 (4.14)
PCG depression score, mean (SD)	2.49 (3.66)	2.42 (3.58)
PCG sex, *n* (%)		
Male	41 (0.4)	161 (1.6)
Female	11,093 (99.6)	9,632 (98.4)
Global health PCG, *n* (%)		
Excellent/very good	7,746 (69.6)	6,760 (69.0)
Good/fair/poor	3,387 (30.4)	3,032 (31.0)
Ethnicity PCG, *n* (%)		
Irish	9,275 (83.3)	8,261(84.4)
White non-Irish	1,203 (10.8)	1,018 (10.4)
Black	295 (2.7)	252 (2.6)
Asian	273 (2.5)	202 (2.1)
Other	53 (0.5)	54 (0.6)

(continued)

Table 1.
(continued)

Characteristic	Total at 9 mo of Age	Total at 3 y of Age
Family social class, *n* (%)		
Professional/managerial	5,340 (48.0)	4,553 (46.5)
Other nonmanual/skilled manual	3,643 (32.7)	3,233 (33.0)
Semiskilled/unskilled	1,148 (10.3)	1,061 (10.8)
Unclassified	1,002 (9.0)	947 (9.7)
Education PCG, *n* (%)		
Lower secondary or less	1,955 (17.6)	1,361 (13.9)
Upper secondary	2,806 (25.2)	3,192 (32.6)
Nondegree	3,112 (28.0)	2,080 (21.2)
Third level	3,249 (29.2)	3,144 (32.1)
Equivalized annual income, mean (SD)	21,507 (13,414)	17,874 (9,551)

PCG, primary caregiver.
[a]Hospital admission (ever): at 9 mo, this question related to an illness or health problem, whereas at 3 y of age, it was for accident or injury.

Classification Tree at 3 y of Age

The CTA for children at 3 y of age is shown in Figure 2. This tree model resulted in a sensitivity of 66% and specificity of 58.5% and overall correctly classified almost 59% of all children in the data set. This data set was partitioned using the CHAID algorithm into 4 levels with 25 nodes, of which 15 were terminal. Ten independent variables were significant in the model; child health (PCG global rating), ethnicity and education of the PCG, history of hospital admission for an injury, family annual income, PCG treated for depression, PCG stress score, total depression score, PCG health rating, and persistence subscale of the LSAC child temperament measure. The most important splitting variable was child global health rating, and of those children who were classified as "not very healthy," 7.1% had a dental problem, whereas of those who were classified as "very healthy," only 4.3% had a dental problem. Approximately 75% (n = 7,274) of the total sample was classified as "very healthy," and the next variable splitter at this node was PCG ethnicity. In the subset of children for whom the PCG was "other white background" (non-Irish white), the prevalence of dental problems was 8.4%, whereas those of children with a PCG of Irish and "any other" ethnicity had a dental problem prevalence of 3.8%. A detailed description of the tree output is contained in Appendix 2.

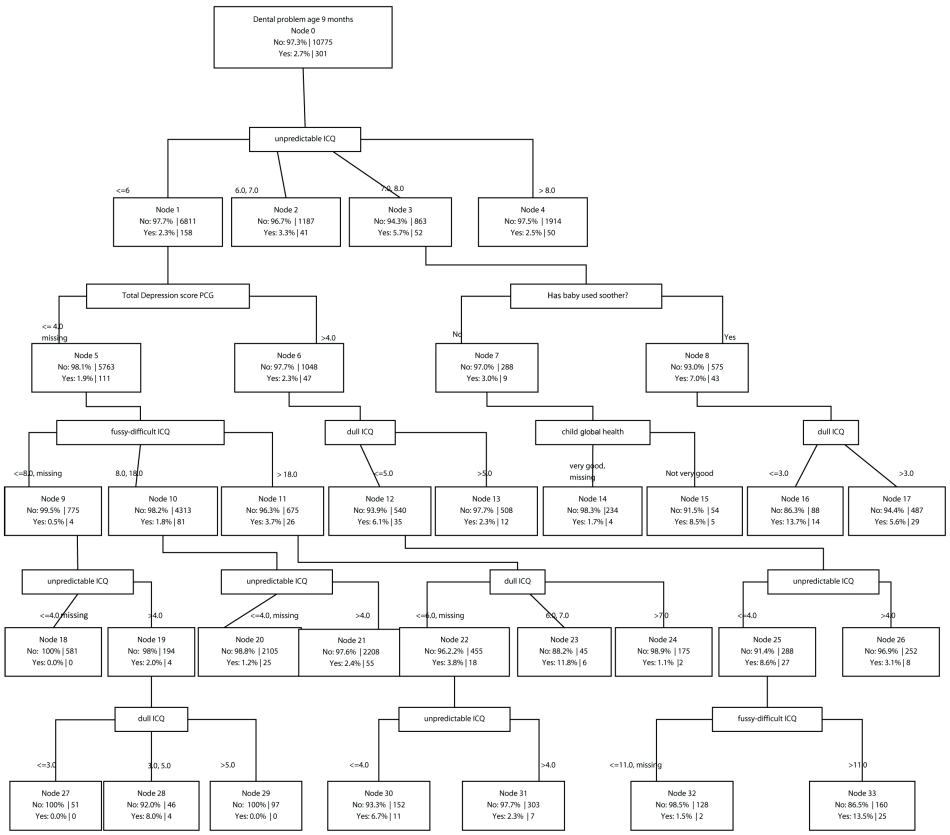
In logistic regression analyses, 5 factors were significantly associated with the likelihood of being a child at 3 y of age with a reported dental problem: child global rating of health (B = 0.44, SE = .11, $P < 0.001$), hospital admission for injury (B = –0.45, SE = 0.12, $P < 0.001$), PCG ethnicity ($P < 0.001$) and education level ($P = 0.01$), and family social class ($P = 0.03$).

Discussion

The findings highlight the relatively high prevalence of reported dental problems in early childhood, with 1 in 50 children at 9 mo and 1 in 20 at 3 y of age with a reported dental problem that resulted in the PCG seeking child health care. Only 16 children had a dental problem at both time points. CTA showed that certain psychosocial characteristics, sociodemographic factors, and measures of PCG and child health were key predictors of dental problems in preschool children. Specifically, infant temperament (unpredictable) at 9 mo and child health at 3 y were the most significant predictors of dental problems. While logistic regression analyses largely supported these findings, the CTA more clearly illustrated the complex multilevel interactions among a greater number of predictors.

This study provided cross-sectional exploratory analyses of the GUI infant cohort (at age 9 mo and age 3 y), which is the largest child population study in Ireland covering key aspects of child health and development. Currently, national data are limited on child dental health, and preschool children were not included in previous national dental surveys (Sagheri et al. 2013). To our knowledge, no study has investigated the association between the prevalence of dental problems in preschoolers and health and psychosocial characteristics of the PCG and child using CTA. On a national population basis, this represented approximately 2,000 infants age 9 mo and 3,200 children age 3 y. This was similar to trends reported in other countries (Slack-Smith 2003; Declerck et al. 2008). However, as our secondary analysis used a PCG-reported dental problem as opposed to, for example, clinically diagnosed ECC, comparisons with other oral health indicators must be made with caution. Prevalence rates for ECC in

Figure 1. Prevalence of reported dental problems at 9 mo of age among classification tree subgroups, percentage, and number in each class. ICQ, Infant Characteristics Questionnaire; PCG, primary caregiver.

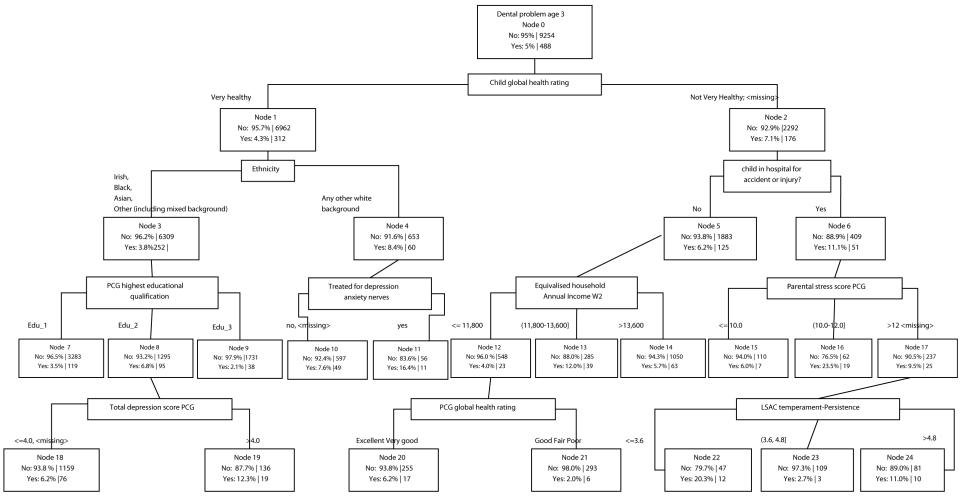


2- to 5-y-olds from other European countries vary from approximately 6% to 28% (Baggio et al. 2015). The actual prevalence of dental problems in our cohort is possibly much higher given that only dental problems for which care was sought were recorded but not the actual outcomes of the visit. It is widely acknowledged that early childhood dental problems are often neglected or untreated unless symptomatic (Slade 2001; Sheiham 2006).

The classification and assessment of important psychosocial variables is a common task in analyzing data from large cohort studies related to child development. Typical statistical approaches include a univariate analysis and construction of a global model using a regression technique to see how well, or poorly, the parameters fit the model. The method of recursive partitioning in CTA facilitates visual identification of complex relationships in a large number of variables among subgroups of a population while not requiring the variable form and distribution assumptions necessary for parametric techniques (Lemon et al. 2003). CTA models multilevel interactions while regression methods largely assume that predictor variables act independently (Kuhn et al. 2014).

It is common when dealing with real problems such as medical or dental classification that the class distribution of the data is imbalanced (Rahman and Davis 2013). Our study target variable involved a binary response with 2 classes where the one of interest was underrepresented, which is described as the minority or positive class (Yap

Figure 2. Prevalence of reported dental problems at 3 y of age among classification tree subgroups, percentage, and number in each class. PCG highest education level subgroups—Edu_1, lower secondary, nondegree, postgrad certificate, upper secondary and technical, professional qualification; Edu_2, primary education, master's degree, primary degree, degree and professional qualification; Edu_3, no education, technical or vocational qualification, doctorate, <missing>. LSAC, Longitudinal Study of Australian Children; PCG, primary caregiver.



et al. 2014). Furthermore, the frequency of dental problems in the data set for 9-mo-old infants was less than 5%, which is described, statistically, as a rare event. Boosting and resampling techniques were applied to attempt to address the class imbalance but did not greatly improve the performance of either model. While class imbalance was greater in the 9-mo data set, the main objective was to generate a classification tree of the characteristics affecting dental problems in the population subgroups. Overall accuracy of the model at 9 mo was high (88.8%), but the sensitivity was low (31.2%), whereas the model at 3 y had high sensitivity (66.0%) and accuracy (58.9%).

Social determinants and oral health problems are strongly related (Watt 2007), as is supported by the results at 3 y of age. However, these factors were not significant predictors of infant dental problems at 9 mo of age. Results from the Victorian Child Health and Wellbeing Study (Renzaho and Silva-Sanigorski 2013) similarly found that socioeconomic factors were only significantly related to child oral health status in older children from 4 to 7 y. While the parental factors most commonly investigated that are associated with child oral health are sociodemographic, there is an increasing awareness of the importance of the PCG and child psychological, general health, and behavioral profiles as risk factors for ECC (Abreu et al. 2015). Our results suggest that the experience of dental problems at 9 mo of age appears to be more related to infant temperament, use of a soother, and PCG depression score.

The regression model also identified those infants at 9 mo of age with a fussy-difficult temperament as significantly associated with the prevalence of dental problems. Similar findings were previously reported with regard to increased prevalence of ECC and infants with a difficult temperament (Spitz et al. 2006; Aminabadi et al. 2014). Poor infant feeding habits, including a high-frequency intake of "sugary drinks" and "night feeds," are high risk factors for ECC (Abreu et al. 2015), although waking of the baby at night for feeds was not a significant predictor in our classification.

Previous conceptual models have proposed that mothers with depressive symptoms and high parental stress may contribute to oral health problems in children through a number of pathways, including early cessation of breastfeeding, unhealthy eating practices, poor oral hygiene behavior, and negatively affecting infant temperament (Kim Seow 2012). CTA (Figs. 1, 2) suggested that while PCG depression score was an important predictor (level 1) for dental problems at 9 mo ($P < 0.001$) for 63% of the sample population, it was only a predictor at level 4 of CTA when the infant was 3 y old ($P = 0.05$). Results of regression analysis also found that PCG depression score was significantly associated with dental problems at 9 mo but not at 3 y of age. However, whether the PCG was "treated for depression, anxiety or nerves" was a level 2 predictor ($P = 0.04$), which split the children at 3 y who were from a non-Irish, white ethnic background (node 4). It is important to note that as CHAID relies on contingency tables for calculating significance tests, continuous

or ordinal variables must be coerced into a categorical form by binning. Thus, some continuous predictors were split by the algorithm at different cut points not necessarily related to clinical significance. For example, the CES-D depression score is a screening tool developed to measure depressive symptomology with "emphasis on the affective component, depressed mood" (Radloff 1977).

The classification tree output at 3 y of age (Figure. 2) showed that child global health was the most significant predictor of dental problem prevalence followed, at level 1, by PCG ethnicity and whether the child had a hospital admission for injury. Although child global health was a second level predictor at 9 mo, this was for a relatively small subgroup (node 7). Evidence for an association between a child's general and oral health continues to strengthen (Sheiham 2006), and adverse childhood experiences, including psychosocial issues, are associated with poorer dental health (Bright et al. 2015). Previous national studies have used PCG-reported child health as a valid and "more holistic" proxy measure for child health as defined by the World Health Organization (Shrivastava et al. 2014). The current analysis strengthens the evidence for the association between overall child health and oral health in a nationally representative sample of preschool children.

In the current study, the "very healthy" children were split by PCG ethnicity, and the infants of those from a non-Irish white background had a significantly higher prevalence of dental problems (8.4%), and this group was subdivided according to PCG being treated for depression or anxiety, which had almost double the rate of dental problems again (16.4%). It is not clear why this particular ethnic group appears to be more at risk for early childhood dental problems, but it may be due to cultural differences related to caries risk factors such as diet, oral hygiene behaviors, or PCG oral health beliefs (Kim Seow 2012).

While PCG education level was the predictor that split the remainder of the "very healthy" subgroup, it is difficult to interpret the results in a meaningful way. However, it is not surprising that PCG education and ethnicity were important components of the classification tree and regression analysis given the strong correlation between these sociodemographic factors and early childhood oral health (Hooley et al. 2012; Abreu et al. 2015).

Results from the logistic regression analysis largely supported the results from the CTA in that child health, PCG education, ethnicity, household income, and family social class were significant factors ($P < 0.05$) associated with reported dental problems at 3 y of age.

It is interesting to note that only 16 of the infants who had a dental problem at 9 mo also had a dental problem at 3 y of age. While the nature of the dental problem may differ at these ages, the associated predictors also varied with sociodemographic factors more relevant at the older child age. This underlines the importance of a life course approach to investigating dental problems and using an exploratory analytical approach that can detect multilevel interactions (Ben-Shlomo and Kuh 2002).

Strengths and Limitations

The GUI study is representative of the Irish population, and a key strength is the range of detailed data collected that are useful in exploring predictors of dental problems in preschoolers. As far as we are aware, this is the first nationally representative study of this age cohort to investigate dental problems and psychosocial factors using CTA. The use of a graphical display tree allows for easier visualization of potentially important interactions and subgroups that might not be discovered using a more traditional statistical approach. Furthermore, as a nonparametric technique, it is not restricted by the form and distribution of the variables being explored and does not require data transformations to use heavily skewed data.

It is important to acknowledge that although the classification tree method is useful from the clinician's perspective, there are a number of limitations such as underfitting, overfitting, and instability, and the interpretation of the tree must be carried out with a degree of critical awareness of what may constitute a plausible relationship. The "oversensitivity" of CTA can cause small changes in input data to result in large changes to the tree appearance as all node splits are dependent on the preceding splits (Kuhn et al. 2014).

The data in the study are PCG reported, and consequently, there is an increased risk of recall bias and social desirability. Although the psychosocial variables used are validated, the study could be strengthened by including information relating to the children's dental condition. The outcome variable measured was PCG-reported dental problem requiring a health care or dentist visit, and this can encompass all oral health problems (including physiological problems such as teething), dental trauma, and early childhood caries. Furthermore, the actual dental problems at 9 mo of age can differ from those at 3 y of age when the child has a more developed primary dentition and risk factors for ECC are more likely to have an impact.

Conclusions

This study provides a clear visual representation, using classification tree analysis, of how PCG and child psychosocial and general health factors are associated with early childhood dental problems, even before the primary dentition is complete. The findings extend previous research by highlighting the relative importance of some known predictors and recognizing the interconnected role of these factors in adopting an integrated multidisciplinary approach to assist in formulating a coherent oral health policy. CTA appears to be a useful, flexible, and appropriate statistical approach to analysis of variable interactions in a large population-based research study without requiring particular distributional assumptions.

Future research should focus on a life course approach to understand the multiple pathways through which these health and psychosocial factors in early childhood may affect oral health throughout life.

Author Contributions

M. Crowe, contributed to data acquisition, analysis, and interpretation, drafted the manuscript; A. O'Sullivan, O. Cassetti, L. Swords, contributed to data analysis and interpretation, critically revised the manuscript; C. McGrath, contributed to conception and design, critically revised the manuscript; M. O'Sullivan, contributed to data acquisition and interpretation, critically revised the manuscript. All authors gave final approval and agree to be accountable for all aspects of the work.

Acknowledgments

Data have been collected under the Statistics Act, 1993, of the Central Statistics Office. The GUI survey has been designed and carried out by the joint ESRI-TCD Growing Up in Ireland study team. The authors received no financial support and declare no potential conflicts of interest with respect to the authorship and/or publication of this article.

References

Abreu LG, Elyasi M, Badri P, Paiva SM, Flores-Mir C, Amin M. 2015. Factors associated with the development of dental caries in children and adolescents in studies employing the life course approach: a systematic review. Eur J Oral Sci. 2015 Aug 14. [Epub ahead of print]

Aminabadi NA, Ghoreishizadeh A, Ghoreishizadeh M, Oskouei SG, Ghojazadeh M. 2014. Can child temperament be related to early childhood caries? Caries Res. 48(1):3–12.

Baggio S, Abarca M, Bodenmann P, Gehri M, Madrid C. 2015. Early childhood caries in Switzerland: a marker of social inequalities. BMC Oral Health. 15:82.

Bates JE, Freeland CAB, Lounsbury ML. 1979. Measurement of infant difficultness. Child Dev. 50(3):794–803.

Ben-Shlomo Y, Kuh D. 2002. A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. Int J Epidemiol. 31(2):285–293.

Berry JO, Jones WH. 1995. The parental stress scale: Initial psychometric evidence. J Soc Pers Relations. 12(3):463–472.

Bourgeois DM, Llodra JC. 2014. Global burden of dental condition among children in nine countries participating in an international oral health promotion programme, 2012–2013. Int Dent J. 64(Suppl 2):27–34.

Bright MA, Alford SM, Hinojosa MS, Knapp C, Fernandez-Baca DE. 2015. Adverse childhood experiences and dental health in children and adolescents. Community Dent Oral Epidemiol. 43(3):193–199.

Condon JT, Corkindale CJ. 1998. The assessment of parent-to-infant attachment: development of a self-report questionnaire instrument. J Reprod Infant Psychol. 16(1):57–76.

Declerck D, Leroy R, Martens L, Lesaffre E, Garcia-Zattera MJ, Broucke SV, Debyser M, Hoppenbrouwers K. 2008. Factors associated with prevalence and severity of caries experience in preschool children. Community Dent Oral Epidemiol. 36(2):168–178.

Dye BA, Arevalo O, Vargas CM. 2010. Trends in paediatric dental caries by poverty status in the United States, 1988–1994 and 1999–2004. Int J Paediatr Dent. 20(2):132–143.

Fisher-Owens SA, Gansky SA, Platt LJ, Weintraub JA, Soobader MJ, Bramlett MD, Newacheck PW. 2007. Influences on children's oral health: a conceptual model. Pediatrics. 120(3):e510–e520.

Goodman R. 1997. The strengths and difficulties questionnaire: a research note. J Child Psychol Psychiatry. 38(5):581–586.

Hooley M, Skouteris H, Boganin C, Satur J, Kilpatrick N. 2012. Parental influence and the development of dental caries in children aged 0–6 years: a systematic review of the literature. J Dent. 40(11):873–885.

Kass GV. 1980. An exploratory technique for investigating large quantities of categorical data. Appl Stat. 29(2):119–127.

Kim Seow W. 2012. Environmental, maternal, and child factors which contribute to early childhood caries: a unifying conceptual model. Int J Paediatr Dent. 22(3):157–168.

Kuhn L, Page K, Ward J, Worrall-Carter L. 2014. The process and utility of classification and regression tree methodology in nursing research. J Adv Nurs. 70(6):1276–1286.

LaVaille PS, Glaros A, Bohaty B, McCunniff M. 2000. The effect of parental stress on the oral health of children. J Clin Psychol Med Settings. 7(4):197–201.

Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. 2003. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Ann Behav Med. 26(3):172–181.

Loh W-Y. 2014. Fifty years of classification and regression trees. Int Stat Rev. 82(3):329–348.

Mäntymaa M, Puura K, Luoma I, Salmelin RK, Tamminen T. 2006. Mother's early perception of her infant's difficult temperament, parenting stress and early mother-infant interaction. Nord J Psychiatry. 60(5):379–386.

Menon I, Nagarajappa R, Ramesh G, Tak M. 2013. Parental stress as a predictor of early childhood caries among preschool children in India. Int J Paediatr Dent. 23(3):160–165.

Moimaz SA, Fadel CB, Lolli LF, Garbin CA, Garbin AJ, Saliba NA. 2014. Social aspects of dental caries in the context of mother-child pairs. J Appl Oral Sci. 22(1):73–78.

Newton JT, Bower EJ. 2005. The social determinants of oral health: new approaches to conceptualizing and researching complex causal networks. Community Dent Oral Epidemiol. 33(1):25–34.

Prior MR, Sanson A, Smart D, Oberklaid F. 2000. Pathways from infancy to adolescence: Australian temperament project 1983–2000. Australian Institute of Family Studies Melbourne [cited 2016 May 6]. Available from: https://aifs.gov.au/publications/pathways-infancy-adolescence.

Quinonez R, Santos R, Wilson S, Cross H. 2001. The relationship between child temperament and early childhood caries. Pediatr Dent. 23(1):5–10.

Radloff LS. 1977. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas. 1(3):385–401.

Rahman MM, Davis D. 2013. Addressing the class imbalance problem in medical datasets. Int J Mach Learn Comput. 3(2):224–228.

Renzaho A, Silva-Sanigorski A. 2013. The importance of family functioning, mental health and social and emotional well-being on child oral health. Child Care Health Dev. 40(4):543–552.

Sagheri D, McLoughlin J, Nunn JH. 2013. Dental caries experience and barriers to care in young children with disabilities in Ireland. Quintessence Int. 44(2):159–169.

Sheiham A. 2006. Dental caries affects body weight, growth and quality of life in pre-school children. Br Dent J. 201(10):625–626.

Shrivastava A, Murtin C, Kelleher CC. 2014. Preschoolers' parent-rated health disparities are strongly associated with measures of adiposity in the lifeways cohort study children. BMJ Open. 4(7):e005328.

Slack-Smith L. 2003. Dental visits by Australian preschool children. J Paediatr Child Health. 39(6):442–445.

Slade GD. 2001. Epidemiology of dental pain and dental caries among children and adolescents. Community Dent Health. 18(4):219–227.

Spitz AS, Weber-Gasparoni K, Kanellis MJ, Qian F. 2006. Child temperament and risk factors for early childhood caries. J Dent Child. 73(2):98–104.

Tang C, Quinonez RB, Hallett K, Lee JY, Kenneth Whitt J. 2005. Examining the association between parenting stress and the development of early childhood caries. Community Dent Oral Epidemiol. 33(6):454–460.

Wake M, Hardy P, Sawyer MG, Carlin JB. 2008. Comorbidities of overweight/obesity in Australian preschoolers: a cross-sectional population study. Arch Dis Child. 93(6):502–507.

Watt RG. 2007. From victim blaming to upstream action: tackling the social determinants of oral health inequalities. Community Dent Oral Epidemiol. 35(1):1–11.

Williams J, Murray A, McCrory C, McNally S. 2013. Development from birth to three years: growing up in Ireland. Dublin (Ireland): Office of the Minister for Children and Youth Affairs [cited 2016 May 6]. Available from: http://www.growingup.ie/fileadmin/user_upload/documents/Second_Infant_Cohort_Reports/BS_Development_from_Birth_.

Yap BW, Rani KA, Rahman HAA, Fong S, Khairudin Z, Abdullah NN. 2014. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. Paper presented at: Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013); Kuala Lumpur, Malaysia.

Proceedings of the Nutrition Society (2016), **75 (OCE3)**, E188 doi:10.1017/S0029665116002032

Summer Meeting, 11–14 July 2016, New technology in nutrition research and practice

Data mapping to augment dietary intake values from a nutritional database to a national cohort survey: protocols to improve quality of reported food intake

M. Crowe*[1], M. O'Sullivan[1], O. Cassetti[1], C. McGrath[2] and A. O' Sullivan[3]
[1]*Dublin Dental University Hospital, Trinity College Dublin,* [2]*Faculty of Dentistry, Hong Kong University and* [3]*Institute of Food and Health, University College Dublin*

The Growing Up in Ireland (GUI) infant cohort was investigated to determine associations between dietary intake, dental problems and anthropometric measurements. Dietary intake is commonly measured using Food Frequency Questionnaires (FFQ). However the FFQ from the GUI survey provided a limited list of "healthy" and "unhealthy foods" and was also limited in terms of the consumption frequency recorded. This study explored the augmentation of this limited dietary intake by unidirectional mapping from the more detailed National Preschool Nutrition Survey (NPNS) which used a 4 day weighed food diary [(1)]. The objectives were: 1) to assess the degree to which the databases could be matched, 2) quantify the dietary food intake that was not covered by the FFQ in GUI for a similar population in terms of frequency and amount of consumption and 3) assess the lack of capture of potentially cariogenic or obesogenic foods by using a short FFQ to determine associations with health outcomes such as dental problems or anthropometric status. Data were derived from the second wave of the GUI infant cohort (n = 9,793) and NPNS (n = 125) of 3 year old children, both sampled in 2010/2011. All of the GUI FFQ categories were mapped in one direction only with food groups from NPNS. Other variables from NPNS such as BMI (Body Mass Index) classification[(2)], social class and gender were also included. The population samples were assessed for difference using the two proportion z- test, $p < 0{\cdot}05$.

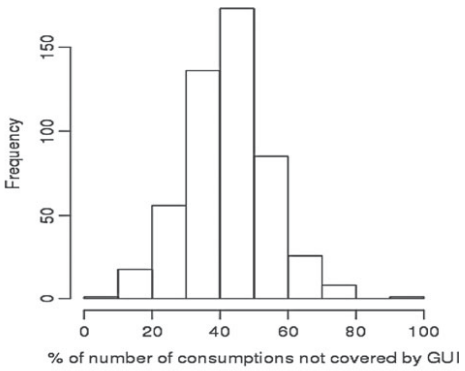


Fig. 1. Number not covered by GUI

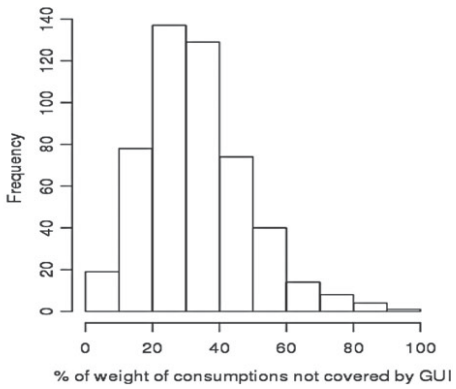


Fig. 2. Weight not covered by GUI

Both samples were nationally representative although NPNS had a higher proportion from the 'Professional' social class. The frequency of consumption of all mapped food groups varied considerably. Approximately 43 % (SD = 13) of the food consumption occasions and approximately 33 % (SD = 15) of the weight of food consumed in the NPNS survey were not covered by the GUI FFO categories. Potentially cariogenic foods such as fruit juices, smoothies, RTEBC, puddings and added sugar were not covered by the GUI FFQ and, on average, each of these constituted approximately 6 % frequency of consumptions overall. Using a limited FFQ in large national surveys restricts the investigation of potentially associated health outcomes. This research demonstrates that if a FFQ provides insufficient dietary information for health outcome analysis, it is possible to augment dietary intake values by data-mapping. This provides more detailed information on food consumption patterns with the potential to enhance health outcome analysis.

1 Irish Universities Nutrition Alliance (IUNA) (2012). The National Pre-school Nutrition Survey. www.iuna.net.
2 Cole TJ, Bellizzi MC, Flegal KM *et al.* (2000). *Br. Med. J.*, **320**, 1240 1243