NUTRITIONAL INFORMATICS: MINING SUPERMARKET SALES DATA AS A NUTRITIONAL ASSESSMENT METHOD

by

Kristina Michelle Brinkerhoff

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

The University of Utah

August 2012

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Kristina Michelle Brinkerhoff**

has been approved by the following supervisory committee members:

John F. Hurdle	, Chair	**5/15/2012** Date Approved
Mollie R. Cummins	, Member	**5/17/2012** Date Approved
Per H. Gesteland	, Member	**5/15/2012** Date Approved
Kristine C. Jordan	, Member	**5/15/2012** Date Approved
Catherine J. Staes	, Member	**5/15/2012** Date Approved

and by **Joyce A. Mitchell** , Chair of

the Department of **Biomedical Informatics**

and by Charles A. Wight, Dean of The Graduate School.

ABSTRACT

Many nutritional assessment techniques, including food frequency questionnaires (FFQs) and 24-hour dietary recalls have innate limitations such as expensive protocols, high respondent burden, and self-reporting biases. Supermarket sales data have shown promise as a new, indirect, inexpensive nutritional assessment method in recent studies. The goals of the research in this dissertation were to link nutritional content to supermarket sales data and to determine the relationship between supermarket purchases and traditional nutritional measures through correlation and regression analyses.

Nutritional data was mapped to sales data at the nutrient and food group levels. One year retrospective supermarket sales data, household food inventory data, and FFQ results were then obtained for 50 households recruited for the study. A correlation analysis was completed to compare percentage of food groups purchased over 52 weeks against food groups in the household inventory and in the FFQ results. Additionally, stepwise regression models were created to predict BMI, energy intake, fat intake, and saturated fat intake based on supermarket sales data.

Nutritional content was mapped to 100% of the supermarket sales data at the food group level and at 69% for the nutrient level. The correlation coefficients between the household inventory and sales data over the course of 52 weeks ranged from -0.13 to 0.83 with an average value of 0.23 at week 32, while correlation for the comparison between the FFQ and sales data ranged from -0.17 to 0.47 with an average of 0.23 at 32 weeks.

The regression models to predict BMI, energy intake, fat intake, and saturated fat intake each yielded significant results for several food group purchases from the sales data.

Mapping nutritional content to sales data was successful, given that there are potential strategies to increase the linkage for nutrient data. The correlation results are in line with other studies comparing nutritional assessment methods against each other and the regression models produced many significant food groups that are substantiated by multiple studies. Overall, the work presented gives an excellent starting point for further informatics research into the untapped potential of supermarket sales data as a nutritional assessment method and public health tool.

To God and my family

TABLE OF CONTENTS

ABSTRACT iii

LIST OF FIGURES ix

LIST OF TABLES x

ACKNOWLEDGMENTS xi

1. INTRODUCTION 1

    1.1 Objectives 2
    1.2 Rationale And Significance 2

2. BACKGROUND 5

    2.1 Current State Of Nutritional Assessment Methods 6
      2.1.1 Nutritional Measures 6
      2.1.2 Dietary Assessment Methods 7
      2.1.3 Additional Nutritional Assessment Methods 10
    2.2 Current State Of Technology In The Nutrition Field 11
      2.2.1 Technology For Professionals 11
      2.2.2 Nutrition Technology For The Public 13
      2.2.3 Technology In Nutrition Literature 16
    2.3 Supermarket Data 17
      2.3.1 Market Basket Analysis 17
      2.3.2 Grocery Receipts 18
      2.3.3 Supermarket Sales Data Studies 18
      2.3.4 Supermarket Sales Data Sources 20
    2.4 Aims 22

3. STUDY PROCEDURES AND DATA COLLECTION 24

    3.1 Data Collection 25
    3.2 Institutional Review Board 25
    3.3 Deciding Upon The Form Of Supermarket Sales Data 25
    3.4 Large Intermountain Grocer Collaboration 25

3.5 Exploratory Data Set 26
3.6 Participant Inclusion Criteria 27
3.7 Participant Enrollment 27
3.8 Household Visit 29

4. MAPPING NUTRITIONAL CONTENT TO SALES DATA 32

4.1 Objective 33
4. 2 Methods 33
4.2.1 Study Design 33
4.2.1.1 Departments 34
4.2.1.2 Commodities 34
4.2.1.3 Subcommodities 34
4.2.2 Data Analysis 36
4.3 Results 36
4.4 Discussion 39
4.5 Conclusions 42

5. CORRELATION STUDIES 50

5.1 Objective 51
5.2 Methods 51
5.2.1 Sales data in food group form from mapping 52
5.2.2 Household Inventory Correlation Analysis 53
5.2.3 Food Frequency Questionnaire Correlation Analysis 54
5.3 Results 55
5.4 Discussion 57
5.4.1 Correlation Coefficients Analyzed by Average R Values 57
5.4.2 Correlation Coefficients Analyzed by Food Groups 61

6. LOGISTIC REGRESSION MODELS 73

6.1 Objective 74
6.2 Methods 74
6.3 Results 77
6.4 Discussion 78

7. APPLICATIONS 94

7.1 Introduction 95
7.2 Interventions 95
7.3 Improve Individual Accountability 96
7.4 Nationwide Surveillance 97
7.5 Importance To Informatics 99

APPENDICES

A. MATERNAL AND CHILD QUESTIONNAIRE 102

B. ADULT HARVARD SERVICE FOOD FREQUENCY QUESTIONNAIRE 106

C. CHILD HARVARD SERVICE FOOD FREQUENCY QUESTIONNAIRE..... 111

REFERENCES 116

LIST OF FIGURES

Figure Page

4.1 Structure of the supermarket database hierarchy 43

4.2 Diagram of mapping nutritional content to supermarket sales data 44

4.3 Distribution of items purchased in total data set.....45

4.4 Distribution of unique food items from the large unmapped population.....46

4.5 Distribution of large unmapped population 47

5.1 Correlation of household inventory and sales data 69

5.2 Correlation of FFQ and sales data 70

5.3 Variation of food group purchases for all 50 households combined over 52 weeks from March 2007 to February 2008.....71

6.1 Histogram of BMI among mothers in 50 households. 86

6.2 Histogram of fat recommendations among mothers in 50 households.....87

LIST OF TABLES

Table Page

4.1 List of food groups used in USDA Nutrient Database 48

4.2 Results of mapping nutrient data to sales data.....49

5.1 Demographic characteristics of the 50 households.....72

6.1 Weight status categories based on BMI.....88

6.2 Logistic regression ORs for BMI.....89

6.3 Logistic regression ORs for energy 90

6.4 Logistic regression ORs for fat recommendations.....91

6.5 Logistic regression ORs for saturated fat.....92

6.6 Food sources of energy among US adults.....93

ACKNOWLEDGMENTS

I would like to thank the members of my committee, John Hurdle, Mollie Cummins, Per Gesteland, Kristine Jordan, and Catherine Staes, for their continual guidance, support, and patience throughout my education. I would specifically like to thank Mollie for her replies to my last minute statistics and data mining questions, Per for his optimism and enthusiasm in my work, Kris for her never-ending encouragement, kind words, and nutrition expertise, and Catherine for her public health guidance and willingness to help with poster and presentation details. I would also like to thank John for being my constant cheerleader, for believing in me and my work, and for the countless hours he put into brainstorming, discussions, and revisions.

The Department of Biomedical Informatics deserves my praise and gratitude, as I have been supported in numerous ways throughout my education by many faculty and staff members, including Joyce Mitchell, Kathy Sward, Scott Narus, Homer Warner, Kathy Stoker, and Jo Ann Thompson. Additionally, Dr. Ed Clark and the National Children's Study (NCS) interest group have been invaluable. Dr. Clark provided the basic idea behind this research, along with the experience and dedication to help me accomplish it. Without his collaboration, this research would not have been possible. The NCS interest group members gave much support, encouragement, and useful suggestions throughout the life of this research.

I have been greatly blessed with multiple funding sources throughout my education. I would like to thank the National Library of Medicine and Robert Wood Johnson Foundation for providing training grants that specifically supported my graduate education. Additionally, the Department of Pediatrics supplied me with an Innovative Grant that allowed this research to take place.

Numerous fellow students provided invaluable support during my graduate work. I want to express my gratitude to Marianne Graf for her time put into organizing the household visits, scanning food items, and data entry of the demographic data; Yuling Jiang for her assistance in data entry for the household food inventories; Phil Brewster for providing knowledge and support and for continuing with the research; and to the many other students who helped in countless ways by providing answers to questions, encouragements where needed, and a genuine interest in my work.

I would like to sincerely thank my family who has helped me in any way possible, even when they did not understand what biomedical informatics was. Specifically, I want to thank my husband, who has supported me in every decision and was willing to listen to me talk nonstop about nutrition, subcommodities, and correlation coefficients.

Lastly, I would like to thank my Heavenly Father for providing me with such an amazing educational opportunity and for sustaining me through the difficult times. This would not have been possible without help from above.

CHAPTER 1

INTRODUCTION

1.1 Objectives

The purpose of this dissertation is to study the feasibility of using supermarket sales data as a nutritional assessment method for individuals and households.

This dissertation is organized into six main chapters:

1. Background
2. Methods and Data Collection
3. Mapping Nutritional Content to Sales Data
4. Correlation Studies
5. Logistic Regression Studies
6. Discussion

1.2 Rationale and Significance

The prevalence of overweight or obese children and adults has increased dramatically in the United States during the past several decades. In 2007-2008 just over one third of U.S. adults were obese and the most recent data for children in 2003-2004 found that 17.1% of children were overweight (1, 2). Although the prevalence has not increased significantly in recent years, the number of overweight and obese adults remains dangerously high and the number of morbidly obese individuals continues to increase each year (3, 4). In addition, it is estimated that there are over one billion overweight adults worldwide, with 300 million classified as obese (5). In urban areas of China, the rate of obesity has risen a shocking amount in preschool children: from 1.5% to 12.6% between 1989 and 1997, while in Great Britain the obesity rates among adults nearly tripled from 1980 to 2002 (6, 7).

These numbers are alarming since obesity is associated with many of the leading causes of death in the United States, including heart disease, diabetes, and malignant neoplasms (3, 8). A study that investigated the underlying factors related to the leading causes of death found that tobacco use, poor diet and physical activity, and alcohol use were the top three actual causes of death in the United States, with 18.1% of deaths attributed to tobacco use, 16.6% due to poor diet and physical activity, and 3.5% resulting from alcohol consumption (9).

In addition to the many health consequences of obesity, there are also numerous economic consequences of obesity. It is estimated that in 2008 approximately $147 billion dollars were spent nationally on medical expenses associated with obesity, almost double the amount spent just 10 years earlier in 1998 (10). The obese also have significantly more hospitalizations, use more prescription drugs, file more professional claims, require more outpatient visits, and incur higher medical costs than the non-obese (11). Along with the negative effects of direct medical costs, obesity has also had an indirect effect on the economy through lowered job productivity and absenteeism (12).

As the prevalence of obesity has escalated over the years, health care professionals have increasingly exerted efforts to combat obesity and the negative health outcomes that are associated with it. These professionals, which include clinicians, registered dietitians, and epidemiologists, use tools to help gain insights into the prevalence, incidence, and trends of obesity along with helping to determine how to fight the increasing prevalence. Some of these tools include studying the nutritional behavior of individuals through questionnaires or medical observation, developing surveillance systems for continual monitoring of populations, such as the National Health and

Nutrition Examination Survey (NHANES), and performing interventions to alter the increase in obesity prevalence.

However, many of the tools used to monitor the nutritional behavior of individuals or populations rely on self-reported data. Self-reported behavior has been shown to have questionable accuracy, with many respondents reporting engaging in healthier behavior than is actually the case (13-15). In an effort to minimize or eliminate inaccurate self-reported behavior, researchers have employed other methods, such as direct observation or obtaining biochemical samples, to determine nutritional status. Unfortunately, these methods are very time-consuming, are associated with a large respondent burden and financial costs, and are generally too resource intensive to employ in studies with a large sample size (16).

The research presented here takes a unique approach to monitoring nutritional behavior through examination of purchasing patterns from a supermarket. The analysis and results demonstrate that sales data have promise within the nutritional, public health, informatics, and medical fields. Using sales data as a nutritional measure has the potential to be a useful method for indirectly monitoring the specific nutritional behavior of individuals, studying entire populations as a surveillance tool, and determining the effects of widespread nutritional interventions.

CHAPTER 2

BACKGROUND

2.1 Current State of Nutritional Assessment Methods

As the knowledge of diet-related diseases has increased and researchers have learned how to best alter the nutritional state to affect health, nutritional assessment has become increasingly important. Accurate and validated assessment methods are needed to obtain precise measures, build surveillance systems with reliable data, and conduct successful interventions. These nutritional assessment methods are based upon four types of measures: anthropometric, biochemical, clinical, and dietary.

2.1.1 Nutritional Measures

Anthropometric data are objective measures about the physical dimensions of the human body and are often used to assess growth in the young, evaluate the health of an individual based on reference standards, or to compare present measurements with past measurements. Common anthropometric measures employed in the nutritional field are height, body weight, waist or head circumference, body mass index, and body fat percentage.

Biochemical methods are another category of nutritional assessment techniques. Biochemical methods involve measuring components of blood, feces, urine, or other tissues of the body that have a relationship to nutritional status. A common biochemical measure is serum cholesterol. Biochemical tests can often detect a change in nutritional status long before anthropometric or clinical signs appear.

Clinical methods include a medical history and physical examination and can be used by clinicians to provide an efficient and timely way of diagnosing and subsequently treating the signs and symptoms of malnutrition without postponing treatment while waiting on the results of biochemical tests. Clinical assessment can be used to diagnose a

variety of conditions including protein-energy malnutrition, HIV wasting syndrome, anorexia nervosa, and bulimia nervosa.

Dietary methods are likely the most commonly used assessment techniques to study nutritional health on a large scale. This method assesses nutritional status by examining food and beverage intake or food consumption patterns over the course of one day up to several months. Dietary methods are widely used by registered dietitians to evaluate individual nutritional intake or to be employed in large, nationwide monitoring surveys to examine the eating habits of populations. Common methods of monitoring include dietary recalls, food diaries, food frequency questionnaires, household inventories, and national surveys.

2.1.2 Dietary Assessment Methods

When measuring diet, there are many different validated methods that researchers utilize. Each method measures a specific aspect of diet and has its own strengths and weaknesses. Researchers measuring dietary intake must take into account the desired aspect of nutrition to measure as well as the limitations of each method while balancing the available resources to choose the most appropriate assessment tool. Several commonly used methods are described below.

The 24-hour recall is an interview administered survey in which the interviewer prompts the respondent to recall all food and drink items consumed over the past 24-hours. The specific nutrient information for each item consumed is then determined from a food composition table or through the use of a computer program to determine the dietary intake of the individual over the 24-hour period. Benefits of the recall include obtaining very detailed information about dietary intake, requiring only short-term

memory, and experiencing relatively low respondent burden (16). The 24-hour recall also does not cause a change in dietary behavior, since it is asking about past behavior (16). Some disadvantages of the 24-hour recall are the expense of hiring a trained interviewer and the tendency of the respondent to report their diet incorrectly due to memory or embarrassment about food consumption (17). Another disadvantage of the 24-hour recall is that it is not necessarily reflective of the respondent's normal diet. Multiple recalls throughout the year must be performed to get a reasonable estimate of usual dietary intake (17).

Food frequency questionnaires (FFQ) are questionnaires that ask about retrospective intake of food items, usually ranging from one month to a year in the past. FFQs are often used because they are self-administered and inexpensive. FFQs ask about intake over an extended period of time, which results in a more representative intake than a 24-hour recall. However, FFQs are subject to self-reported errors and the memory of the respondent. Also, many FFQs do not ask respondents about portion size and instead calculate standard portion sizes based on the respondent's sex and age, which could provide inaccurate data.

There are many commonly used validated food frequency questionnaires, including the Diet History Questionnaire (DHQ) developed by the National Cancer Institute, the Willett and Block questionnaires named for the individuals who helped develop and validate the tools, and the Harvard Service Food Frequency Questionnaire (HSFFQ) developed at the Harvard School of Public Health.

While the goal of all FFQs is to accurately measure dietary intake, there are marked differences between questionnaires. Some of these differences include the length

of the questionnaire, available languages, age groups offered, analysis options, and detail of the questions asked. For example, the HSFFQ is a self-administered four page questionnaire offered in both English and Spanish that requires only basic literacy, has validated forms for women and children aged 1 to 5 years, can be paper or computer-based, and assumes portion sizes based on age and gender (18, 19). In contrast, the DHQ used in NHANES has four different formats, which are all available in a paper or Web based version, is approximately 40 pages long, includes 134 food items and 8 dietary supplement questions, and asks questions about frequency of consumption along with portion size (20). With the wide variety of questionnaires available, researchers using FFQs must evaluate the benefits and limitations of each questionnaire to determine which one is most appropriate for their work.

A food diary is another commonly used dietary measure. A food diary requires the participant to manually record all food items and portion sizes of the foods consumed over a specific time period, typically ranging between 1 and 7 days. Advantages of a food diary include obtaining a detailed description of intake over several days and acquiring nutritional intake data that does not rely on the respondent's memory. Some disadvantages include the possibility of behavior change due to observation or the complexity of weighing and recording foods. In addition, the food diary does not have high respondent rates since it requires very cooperative participants (17, 21).

Analyzing diet by direct observation through photographic or video records is another technique occasionally used in the nutritional field. This method is very exact of current intake, but may not represent usual intake, as the participant might alter his or her behavior while being watched. Also, direct observation has a high respondent burden and

is a very expensive nutritional assessment technique to employ. Although direct observation is not a practical measurement for many studies, it is easier to employ in a closed setting such as a hospital, nursing home, or prison, where dietary intake can be closely monitored (16).

While all of the methods above have been validated and used in numerous nutritional studies, there is not a single best method due to the complexity of measuring dietary intake. Despite the limitations, correctly collecting and analyzing data from a properly chosen method can yield useful and significant results.

2.1.3 Additional Nutritional Assessment Methods

The methods described above all have the goal of estimating the dietary intake of an individual. However, there are other measures of nutrition that are also commonly used in studies to help researchers understand nutritional behaviors. These measures include food availability, food security, and food disappearance.

Food availability, or the amount and types of food in a household, is generally measured through a household food inventory. Household food inventories can take the shape of two forms: an open inventory in which trained researchers visit the participants' homes and manually record all food items present, or a predefined inventory where a checklist of food items is presented to the household to determine whether each item is present or absent (22). The data are often analyzed according to food groups or specific types of food, such as high fat foods or fruits and vegetables present in the household. Although food availability is not a direct substitute for dietary intake, food inventories are used to assess the household food environment since the availability of foods are likely to affect nutritional choices.

Food security measures the access of members of a household to enough food to live an active, healthy life. Food security also includes the availability of safe, nutritious food and being able to obtain that food through socially acceptable ways. Food security is an important measure because it is one of several factors that allows a population to be healthy (23). Food security is usually measured through a household questionnaire, with the results guiding policy and program design (24). Federal programs such as the Food Stamp Program, WIC, and the National School Lunch Program along with Community Food Assistance Programs (food pantries and soup kitchens) rely on food security studies to better serve communities in need (23).

Food disappearance is another method used to indirectly measure the nutritional habits of a population. Food disappearance measures the food consumed by a country or other large population over a period of time. This method is calculated using the food balance sheet, which takes into account numbers on food production, imports, exports, and estimates for spoilage or waste foods. Food disappearance is a useful measure to gauge the nutritional habits of an entire country and can be useful for formulating nutritional policy and programs.

2.2 Current State of Technology in the Nutrition Field

2.2.1 Technology for Professionals

Computerized programs for nutrition professionals were first discussed in the late 1950s, but really flourished in the early 1980s to help with the routine calculations required for their jobs (16). Today, technology in the nutrition field can range from simple handheld calculators that determine measures such as estimated energy

requirement and resting energy expenditure to computerized dietary assessment programs used for calculating detailed nutrient intake from 24-hour recalls.

Computerized dietary assessment programs are systems used by registered dietitians to organize and analyze food recalls, food frequency questionnaires, and recipes. The data, usually given by the respondent, have to be manually entered into the computer or, in the case of some food frequency questionnaires, optically scanned. The user can then search within the program's nutrient database for food items to match the respondent's answers. Since no one nutrient database is complete, dietitians regularly have to find a substitute food item that best estimates the nutrients from the respondent's record. Once the data are completely entered in, the user can print reports or export the data to be analyzed further. Using a computerized dietary system decreases the time, labor, and expense associated with analyzing dietary intake and also has the potential to decrease errors in calculations. Although computerized systems are an improvement over looking up foods in a table and doing nutrient calculations by hand, the systems still require extensive manual labor in searching for the correct food items and can result in skewed data for the food items that are not in the database.

Another common use of technology in the nutrition field is the development of computerized surveys, such as food frequency questionnaires (FFQ). While many FFQs are still completed with paper and pencil, some of the more commonly used questionnaires are now also available as a computer application or over the Internet. The Diet History Questionnaire (DHQ), for example, is a FFQ developed by the National Cancer Institute and is also available in a Web-based version. Computerized surveys provide many benefits, including prompting respondents to complete all questions before

advancing to the next, having automatic skip patterns for questions that do not apply to the respondent, and incorporating an added level of data quality by having algorithms that check for inconsistent or implausible answer (25).

The merger of nutrition, information science, and technology has also yielded products such as applications on personal digital assistants (PDAs) to help with calculations or other nutrition knowledge, small handheld calculators that calculate regularly used diet or energy formulas, and Excel spreadsheets created with formulas embedded in the cells. These applications of technology save dietitians the time and labor of meticulously calculating values like resting energy expenditure or estimated energy requirements, while reducing the opportunity of error in the calculations. Although small calculation errors in energy requirements for most adults would not be life threatening, at-risk groups such as premature infants and burn victims have very specific dietary needs. Implementing computer applications to help with calculations gives an added assurance that energy expenditure or requirement figures are correct.

2.2.2 Nutrition technology for the public

Nutrition is often a difficult subject for the public to understand, as they are frequently left to sort fact from fiction on their own. In addition, the media do not always make matters easier. The results from scientific studies are at times reported only in bits and pieces and sometimes seemingly contradict previous studies (26). Dietary recommendations from organizations such as the United States Department of Agriculture (USDA), the American Heart Association, and the Department of Health and Human Services are also not always well understood by the public and can be difficult to implement into the lives of individuals and families.

Due to the complexity of understanding nutrition, many new resources have emerged with the goal of informing the public and helping them to develop and maintain healthy eating habits. While many of these new resources are Internet-based, there are also other uses of technology, such as computer applications, to help individuals track dietary goals.

One such example of the utilization of technology for the public is the use of PDAs for diabetic patients. Patients are traditionally encouraged to keep a record of their blood glucose levels, since patients who are diligent about monitoring their glucose levels tend to have better glycemic control, which is associated with better health outcomes. While in the past blood glucose levels, along with insulin doses, dietary intake, and physical activity, have been tracked with pencil and paper, the current technology allows patients to have an electronic diary, either on a personal computer or on a portable PDA (27). Certain blood glucose monitors even allow the readings to be electronically transferred to the computer, eliminating the need to manually type the results. Although tracking glucose levels, insulin doses, dietary intake, and physical activity will not appeal to all diabetes patients, the benefits are numerous. In addition to automatically downloading data from some blood glucose monitors, electronically tracking health information allows for increased legibility, the ability to quality check for mistyped information, the power of analysis or trends of data through graphs or tables, and the capability to send the information to a physician for inclusion in the patient's personal health record. Similar technology has also been developed for use by dialysis patients who also need to track their dietary intake (28, 29).

Another common application of technology in the nutrition field is found through resources on the Internet. There is an ever increasing supply of diet and health information on the Internet, with the data on many Websites of questionable reliability. However, there are many dependable Websites that not only have accurate data for the public, but also include a variety of interactive features to encourage healthy behaviors, such as dietary intake monitors, menu planning, and exercise tracking.

The Choose My Plate Website (choosemyplate.gov), created by the USDA, offers many services for the public based around recommended intake of food groups. Included in the Website are: tips on how to understand and use the new MyPlate graphic as a nutritional tool; the SuperTracker to monitor food intake and physical activity for an "in-depth assessment of your diet quality and physical activity status;" the Menu Planner to help plan menu choices based on the food pyramid's guidelines; and podcasts to discuss and demonstrate suggested dietary behavior changes (30).

The popularity of smart phones, iPods, iPads, and similar devices has created a new venue for electronic nutritional applications. Applications like *MyFitnessPal's Calorie Counter* and *Diet Tracker* are abundant and can be an excellent instrument in helping individuals track their daily caloric intake, exercise, and weight loss. *Glucose Buddy* and *Carb Counting with Lenny* both provide a means for managing diabetes by tracking carbohydrate consumption, glucose levels, and exercise on handheld devices.

There are many other Websites and electronic tools using modern technology in similar ways to motivate and encourage individuals who are striving to live healthy lifestyles. While these Websites are not a one-size-fits-all dietary solution, they are prime

examples of using technology to distribute accurate, helpful nutritional information for those searching for it.

2.2.3 Technology in Nutrition Literature

As technological advances in the nutrition field continue to grow, technology reported in recent nutritional literature similarly increases. Although the household food inventory, described in Section 2.1.2 above, provides useful data on food availability, it is rarely administered due to the immense costs and excessive respondent burden involved. Hoping to decrease the cost and respondent burden of household inventories, Weinstein et al. conducted a study using a Universal Product Code (UPC) scanner to record food items in a household (31). The scanner was found to be almost as accurate as a traditional line-item inventory using pencil and paper and took 31.8% less time. The study concluded that the UPC scanner was a feasible tool for household food inventories. Combining electronic devices, such as a UPC scanner, with nutritional knowledge has the potential to diminish some of the traditional limitations of time-intensive assessment techniques (31).

Another study tested the feasibility of using smart card technology to record lunch choices of children in a school cafeteria (32). The smart card system was created to allow an electronic audit of individual dietary choices that was linked to a food composition database to determine the nutritional profile of each item consumed. The study found that using smart card technology is a feasible way to monitor the eating habits and trends of individual school children or populations over an extended period of time. While this approach clearly measures food choices and not actual intake, the data are still a valuable measure of nutritional behavior within a school setting. The methodology has the

potential to be applied to other closed settings such as prisons and nursing homes to better understand dietary choices of additional populations (32).

2.3 Supermarket Data

2.3.1 Market Basket Analysis

Supermarket data used as a means of nutritional assessment is a relatively new idea in the nutrition field. Traditionally, commercial sales data from supermarkets have used market basket patterns for determining product placement within a store, pricing strategies, and other marketing decisions (33-38).

Market basket patterns often use data mining techniques to look for association rules such as "90% of transactions that purchase bread and butter also purchase milk" (35-37). There have also been a variety of other studies examining information like frequency of basket size, distribution of basket size by hour of the day, and a correlation of trips versus expenditure at the grocery store (35, 38). Market basket analysis is generally used for the economic and marketing purposes of stores, but the principles and variables studied could have meaningful results and implications in the nutrition field.

A study conducted with grocery store purchases in Texas used the data from a marketing research database to analyze associations between dairy purchases and ethnicity over a 13-week period of time (39). Although limited data were available from the marketing database, patterns between purchased dairy products and ethnicity were found, and it was concluded that purchase records, linked with nutritional information, could be used for large-scale epidemiological studies (39).

2.3.2 Grocery Receipts

In addition to market basket studies involving large datasets, there are a few smaller studies that have collected supermarket receipts to study market basket patterns along with demographic information at the individual level (40-42). One study collected receipts, demographic information, and data on eating habits and body-image perception from shoppers (n=48) at supermarkets in Kentucky (41). The data were analyzed to determine the relationship between money spent on high-fat or low-fat foods and eating habits or body image perception. Although this study only collected sales data from receipts for one shopping trip, meaningful and interesting associations were found between food choices and perceived body image. The study also showed that interesting associations can be found without knowing the exact nutritional make-up of the foods purchased. For example, households where no one was noted as overweight spent less money on foods in the fats, oils, and sweets categories than households where at least one individual was perceived to be overweight. By categorizing the food items into similar nutritional categories, notable relationships were found.

2.3.3 Supermarket Sales Data Studies

Along with the supermarket receipt study described above, several recent studies have shown that sales data gathered directly from the supermarket hold promise as a nutritional assessment method for long-term public health surveillance and for the evaluation of nutritional interventions (39, 43-50).

One such study researched strategies to promote healthier purchasing patterns through a randomized intervention (50). Participants were randomized into one of four groups: price discounts on predefined healthy foods, nutrition education, price discounts

and education, and no intervention. The study was able to successfully track purchases from the participants over a 12-month period and found that price discounts increased the percentage of healthy foods purchased while education made no difference on foods purchased (50). Results from such studies are extremely valuable, as supermarkets and public entities can now be better informed about intervention effects and allocate their resources into programs that will likely positively influence healthy eating behaviors.

In addition to studies completed by researchers, supermarkets themselves are starting to realize the potential of their data from a nutritional standpoint. The supermarket chain Safeway has recently released a program called FoodFlex, which allows customers to track their purchases online (51). Once enrolled in Safeway's free club card program, customers can sign up for a FoodFlex account on Safeway's Website at foodflex.safeway.com. Their account will then be automatically populated with data about their Safeway food purchases made using the club card. The data are available in multiple formats including a line-item list of recent purchases, estimates of household intake compared to USDA guidelines, household nutrition trends over time, and personalized suggestions tailored to dietary needs (51).

The Safeway FoodFlex is a novel program that aims at improving awareness of purchasing patterns and providing suggestions for alternative healthier food items. While it is hypothesized that a program such as FoodFlex will have a positive impact on the nutritional well-being of households, the effect of such a program is not known. Supermarket sales data being used as a nutritional assessment tool is still largely under-researched and methods to increase the ease of access to and use of sales data need to be explored further.

2.3.4 Supermarket Sales Data Sources

There are many potential options for sources of supermarket sales data. As stated above, supermarket receipts gathered from customers have been successfully used to determine relationships between dietary intake and perceived body size (41), education and ethnicity (40), and household food purchase behavior (42). However, there are shortfalls to this method of data collection. First, it can be difficult to obtain receipts from all household food purchases, as individuals and families frequent multiple supermarkets, restaurants, and other establishments for meals. Obtaining a receipt, keeping it, and turning it in to the research team does not always happen for every item consumed (42). Second, the receipts are often annotated by the participant to provide additional information about type of item and serving size. This relies on self-reporting and is known to produce a bias (13-15). Third, keeping and annotating the receipts can result in high respondent burden, leading to noncompliance of study procedures or to attrition (42). Lastly, receipt studies result in high researcher burden due to the time-intensive task of transferring all receipts to an electronic form for management and analysis and to determine the nutritional make-up of each food item.

In other studies, purchasing data are obtained through the use of a home scanning protocol. Households are given a UPC scanner, are asked to scan every food item brought into the home, and have to manually write down the items without a UPC. While this option is very thorough, capturing all food items entering the home, the respondent burden is extremely high and can result in omission of items or fatigue. In addition, respondents might display the Hawthorne effect and alter their behavior because they are

aware that they are being studied (16). The home scanning option is also very expensive, as each household participating must have a scanner provided for them (17).

Another option for obtaining purchasing records is to acquire them from businesses that collect sales data from retail and grocery stores. There are commercial databases owned by companies such as Information Resources, Inc. (IRI) and Nielsen (formerly known as AC-Neilsen) that can be obtained by researchers wanting supermarket scanning data. Databases that include sales data are a useful nutrition tool because they contain objective data without any participant burden. These data are largely used for marketing purposes and while they can be used for nutrition research, it can cost up to hundreds of thousands of dollars for current year data (52). The databases do possess a wealth of information contained in millions of purchasing records; however, the records are not linked to any demographic data, making conclusions relating to specific households or populations very difficult. In addition, the data were not meant for academic research purposes and detailed documentation on sampling and data collection procedures are generally not available (52).

Obtaining supermarket sales data directly from a supermarket would be ideal, as researchers would not only receive an objective measure without any respondent burden, but they could also work with the supermarket and customers to obtain demographic information. However, due to reasons such as privacy, confidentiality, and proprietary concerns, supermarket chains are not eager to release their sales data to researchers.

There is another major barrier to utilizing supermarket records: there is no automatic linkage between the nutritional content of purchases and the data records that represent them in the supermarket databases. The majority of food products manufactured

and sold in the United States have both a Nutrition Facts Label and a UPC printed on their packaging; however, there is no all-encompassing database that links them together. To use sales data effectively for nutritional assessment the UPC must be linked to some type of nutritional data. That could take the form of a detailed listing of macro- and micro-nutrients for each food item, or the configuration of sales data grouped by similar nutritional attributes, such as the classic United States Department of Agriculture (USDA) food groups.

Due to the difficulty in linking nutritional information to sales data, early efforts to study sales data did not obtain nutritional content for every food item within a supermarket database. Instead, the analyses focused on specific but limited subgroups of purchases, such as dairy products or a supermarket's top-selling items (39, 46).

One resource with potential for linking nutrient information to supermarket sales data is the USDA Nutrient Database for Standard Reference (U-NDSR) (53). The database contains over 7,500 coded food items, complete with nutritional information such as calories, carbohydrates, protein, fats, fiber, sugars, minerals, and vitamins. The database is a rich source of food-specific nutritional information, yet there is no way to automatically link the nutritional information to supermarket sales records through the UPC.

2.4 Aims

The aims of the work presented in this dissertation include:

1. Determine the feasibility of utilizing supermarket sales data as an assessment tool.

2. Study the relationship of sales data to traditional nutritional measures of a household.
3. Study the predictability of sales data to nutritional metrics of an individual household member.

CHAPTER 3

STUDY PROCEDURES AND DATA COLLECTION

3.1 Data Collection

To successfully fulfill the study aims presented in Chapter 2, four types of data were required: first, sales data acquired from a supermarket chain; second, dietary intake data from one or more members of the household that are contributing to the sales data; third, household inventory data from the household; and fourth, individual and household demographic data. The following sections will describe the acquisition of these data sources in detail.

3.2 Institutional Review Board

Previous to the attempt to acquire sales data, a study protocol was submitted to the University of Utah Institutional Review Board detailing and the study was found to be exempt.

3.3 Deciding Upon the Form of Supermarket Sales Data

Taking note of the strengths and limitations of the different sources of sales data and knowing that the purposes of this study included linking sales data to individual and household dietary information, it was decided that obtaining supermarket sales data directly from a supermarket would be ideal. As these data are not readily available, we had to recruit the help of a local supermarket chain and enter into an agreement with them.

3.4 Large Intermountain Grocer Collaboration

Performing a study analyzing supermarket sales data required certain criteria from a supermarket chain. The supermarket chain needed to have a frequent shopper program that customers enroll in to keep track of household purchases, be one of the largest

supermarket chains in the Salt Lake Valley to assure that participants would be purchasing a large number of grocery items from the store, be willing to contact customers that could potentially be participants in the study, and be willing, after informed consent from the participating customers, to share the customers purchasing records with investigators of the study.

There are only a couple supermarket chains within the Salt Lake Valley that would fit the criteria listed above. However, a specific Large Intermountain Grocer within the Salt Lake area was our first choice, as they have a long standing relationship with supporting endeavors of the University of Utah and Primary Children's Hospital.

To gauge the interest of the Large Intermountain Grocer in supporting this research work, a member of the research team personally met with the president of the Large Intermountain Grocer. After the president expressed interest in collaborating with the study, we prepared and presented a research proposal to several other employees, including the Vice President of Public Affairs and the database administrator. Upon continued interest, we settled upon a research protocol that worked for both parties and officially formed the collaboration.

3.5 Exploratory Data Set

Once the collaboration was formed, the Large Intermountain Grocer provided a de-identified data set from their database for exploratory analysis in preparation for receiving the participants' data. These data included a random sample of purchases over a period of 2 weeks from approximately 16,000 nonidentifiable customers enrolled in their Frequent Shopper Card program. The food and beverage purchases from this population over the two week time period accounted for over 2 million items purchased.

3.6 Participant Inclusion Criteria

Participants enrolled in the study met the following inclusion criteria:

1. Participant is a Frequent Shopper Card member at the Large Intermountain Grocer with at least 12 months of retrospective purchases linked to their card. Having a complete years worth of data is suggested for nutritional studies to avoid seasonal changes in dietary intake or purchasing behaviors (54).
2. Participant is classified by the Large Intermountain Grocer as one of their "top tier shoppers" ensuring that the participant did shop at specific supermarket chain frequently.
3. Participant resided in the Salt Lake Valley.
4. Participant's household must be the primary dwelling place of a child aged 1 to 5 years old and the biological mother of that child. This criterion was formed to maintain compatibility with the Utah site of the National Children's Study, as this work is potentially a precursor to research for the National Children's Study.

3.7 Participant Enrollment

Participant enrollment followed a rigid protocol to ensure inclusion criteria were met. Data were kept secure and confidential. The Large Intermountain Grocer contacted the Frequent Shopper Card customers through a recruitment letter (see Appendix). The letter was written by the research team, but approved and mailed by the Large Intermountain Grocer at the end of December 2007 to the homes of top tier shoppers within the Salt Lake Valley who had greater than 12 months of retrospective sales data.

Interested participants who met the inclusion criteria could contact the University of Utah research team by phone number for further information and enrollment. If no action was performed by the potential participant, no further contact would be made and researchers would have no access to any of their personal or sales data. As we were collaborating with the Large Intermountain Grocer, it was imperative that we notify potential participants that no data were shared with the research team unless they gave informed consent to participate in the study.

To field the calls of potential participants, a 24-hour answering service was hired to obtain potential participants' information including: first and last name, phone number, address, and Frequent Shopper Card number. The potential participant was then notified that a research representative would contact them shortly for further information on the study. The answering service emailed us contact information for each potential participant as they received it, enabling a representative to quickly contact participants for more information. During the return call from a member of the research team, potential participants were asked to verbally verify that they met all inclusion criteria. Name, address, and Frequent Shopper Card number were gathered again so the Large Intermountain Grocer could double check that the interested customer was in fact the intended recipient of the recruitment letter. Upon verification of meeting inclusion criteria, the participant was put on a list and informed that they would be contacted shortly to schedule a visit date with the research team.

Owing to budget and time constraints, we planned on recruiting 50 households for the study. After mailing the recruitment letters, we received an overwhelming response and stopped taking information from potential participants after over 100 households had

contacted our answering service. Any household contacting the answering service after our enrollment ended was connected to a voice message letting them know that we appreciated their interest, but due to the great response we had received from households, the enrollment was closed.

The first 50 households that successfully met all of the inclusion criteria were chosen for participation in the study. We did not attempt to get a representative sample of household the geographical area, as this was a pilot study to inform future work and the study goals were not to determine information about the population, but rather explore the relationship between different types of nutritional assessment methods.

The study participants were given the appropriate informed consent documents approved by the University of Utah's Institutional Review Board (IRB) at the beginning of the household visit. Upon completion of the visit, the household was given a $50 gift card to the Large Intermountain Grocer for their time. The gift cards were made possible through a grant supporting the study from the University of Utah's Department of Pediatrics.

3.8 Household Visit

The objective of the household visit was to obtain:

1. Signed copies of the informed consent forms.
2. An adult Harvard Service Food Frequency Questionnaire (HSFFQ) completed by the mother in the household.
3. A child HSFFQ completed by the mother for her child aged 1-5 years old.
4. A demographic form completed by the mother.
5. A household inventory detailing all food items found in the household.

Upon arrival at the household, the two research team members first sat down with the mother and explained to her what we would be doing during the visit and what data we would gather from her and her household. We then had her read through the informed consent document and sign it if she was still interested in participating. She then received instructions on how to complete the demographic questionnaire and the HSFFQ for herself and her child. If she had more than one child between 1 and 5 years old, we randomly selected one of the children to be the subject of the child HSFFQ.

As the mother started filling out the documents, the research team began performing the household food inventory. We followed the protocol presented by Weinstein et al. in which a handheld Universal Product Code (UPC) scanner was used to collect barcode data on food and beverage items within the household (31). As the scanner data was to be linked to a UPC database to obtain detailed nutritional content for items in the inventory, photographs of the Nutrition Facts label were taken in anticipation of some food items not appearing in the UPC database. Data for the items that were not included in the UPC database could then be entered after referencing the correct Nutrition Facts photograph. This process allowed for complete data for the household inventory. Upon completion of the inventory, the data would then be loaded onto a computer for data management and analysis.

Food and beverage items without a barcode, such as fruit, vegetables, or homemade dishes, were manually recorded during the visit and later entered into the computer under the appropriate household. The nutritional content of such items were looked up in the USDA nutrient database. Items that were in the house to be used strictly as emergency food storage and were not in the regular rotation of items consumed were

noted but not included as part of the household inventory, as these specific items were not regularly purchased or consumed. These items were frequently stored in sealed metal number 10 cans or large buckets that are capable of holding mass quantities of grains or other foods items.

Upon completion of the household inventory, the surveys were gathered from the mother and checked by a member of the research team to ensure that they were filled out correctly and completely. The mother was then given the $50 gift card for her participation in the study.

CHAPTER 4

MAPPING NUTRITIONAL CONTENT TO SALES DATA

4.1 Objective

The objective of this portion of the study was to use the U-NDSR to assess the feasibility of linking nutritional content from the U-NDSR to supermarket sales data for a large supermarket chain's sales database. Such a linkage would permit a detailed nutritional assessment for all food item purchases made within the supermarket chain. We aimed to map both U-NDSR nutrient information and food groups to supermarket sales data to more fully characterize the usefulness, from a nutritional assessment standpoint, of the commodity organization that supermarkets typically use. If successful, the mapping of both nutrient information and food groups will permit researchers greater versatility in analyzing the nutritional content of sales data.

4. 2 Methods

4.2.1 Study Design

A de-identified data set was acquired from a supermarket chain with stores throughout the Intermountain West. These data consisted of approximately two million purchases from random customer visits made over a two-week period in August 2007. The records were chosen to provide a large variety of items purchased. The large random sample likely, but not necessarily, reflects a representative sample of the geographical area and the supermarket's customer base in summer.

Data fields within our sample data set included: UPC, text description of item, date and time of purchase, some indication of quantity and pricing, and three store-specific hierarchical categories used to organize items within the supermarket and database. The top tier in the hierarchy was called "department;" the next was named

"commodity," and the most granular category was entitled "subcommodity" (Figure 4.1). We describe the organization in more detail in the paragraphs below.

4.2.1.1 Departments

There were nearly 70 departments in the supermarket database, with each containing strictly food items, strictly nonfood items, or a mixture of items. For example, the Fresh Meat, Ice Cream, and Soft Drink departments contained only food items while the Cosmetics, Garden, and Sporting Goods departments contained only non-edible items. There were a few departments such as the Baby department that contained a mixture of food and nonfood items. It would be hard to infer many nutritional details from only the department alone.

4.2.1.2 Commodities

The second tier in the hierarchy, the commodity category, was very similar to the department category. As in the departments, the majority of the commodities were only food or nonfood, but there were occasionally commodities that had a mixture of food and nonfood items. Examples of commodities include Yoghurt (department Dairy), Cookies (department Bakery), and Salad Dressing & Sandwich Spreads (department Grocery-all other). The roughly 400 commodities were more granular than the departments and each commodity only belonged to one department.

4.2.1.3 Subcommodities

The subcommodity category was the most granular grouping within the supermarket database. All of the over 2,100 subcommodities were either classified as

food or nonfood categories and each item in the database was assigned to a subcommodity with similar items. Examples of subcommodities include alarm/clock radios, authentic Thai foods, kiwi fruit, and sour creams.

To obtain detailed nutritional content (calories, carbohydrates, fats, proteins, etc.) for the supermarket data set, unique values of the subcommodity level of the supermarket's classification system were manually mapped to a corresponding food item from the U-NDSR. The subcommodity category was chosen for mapping because it represented the finest level of granularity among the supermarket categories and showed the closest similarities to the level of groupings in the U-NDSR. In addition, a single department or commodity category might include both food and nonfood products, making it difficult to sort out the nutritionally relevant items.

We attempted to automatically map by using a string matching between the subcommodity categories and the food item field in the U-NDSR database; however, there were only four matches out of the 1,252 food related subcommodities from the supermarket data set. The number of matches was low because the two data sets were characterized by different naming strategies. The supermarket's text descriptions were generally short and used many abbreviations, while the U-NDSR food item descriptions were lengthy and used complete words.

Due to the inability to automatically match U-NDSR nutrient information to the supermarket data set, the data were mapped manually. This process involved taking each subcommodity category and searching for the equivalent food item in the U-NDSR (Figure 4.2). The number of successfully mapped subcommodities was recorded. In addition, the number of subcommodities that could not accurately be mapped to a single

U-NDSR food item and the specific barriers impeding the mapping process was documented.

In addition to mapping detailed nutrient information to the supermarket sales data, food groups defined by the USDA (Table 4.1) were also mapped to each subcommodity category. For subcommodities successfully mapped to a U-NDSR food item, the food group was directly imputed from the U-NDSR food item code. The subcommodities that were not successfully mapped to a food item were manually matched to a U-NDSR food group.

4.2.2 Data Analysis

Descriptive statistics were used to determine the success rate of mapping nutrient information to sales data. To address the large number of unmapped subcommodities, we performed a subanalysis to determine the distribution of the food items and to formulate a strategy that would allow nutritional data to be linked to these unmapped items. All statistical analyses were conducted using R (version 2.6.2, 2008, R Foundation for Statistical Computing, Vienna, Austria).

4.3 Results

The 2,009,533 purchased food items in the data set clustered into 29,981 unique UPCs and 26,854 unique items. The distribution skew was substantial (Figure 4.3); some items were purchased often and others very rarely. The supermarket sales database was organized as follows for food specific items (distinct counts in parentheses; counts exclude nonfood items):

Store Departments (N = 36)

Commodities (N = 210)

Subcommodities (N = 1,252)

Items (N = 26,854)

UPCs (N = 29,981)

Purchases (N = 2,009,533)

The results of mapping food items from the U-NDSR database to the supermarket subcommodity categories are shown in Table 4.2. Since the distribution of food items purchased from each subcommodity category varies, the results are reported at the subcommodity level, the unique food item level, and for the entire data set. Approximately 70% of the subcommodities were successfully mapped to a U-NDSR food item and therefore have complete nutritional data linked to them. This equates to a success rate of 69.1% for the purchased items in the entire data set.

In the process of linking U-NDSR nutrient information to the subcommodities, three barriers to successfully mapping the complete data set were found. The most common barrier, accounting for the unsuccessful mapping of 21% of the subcommodities, was the presence of heterogeneous subcommodities containing nutritionally diverse food items that were unable to be mapped to a single U-NDSR entry. For example, the "soft drink" subcommodity category includes regular and diet soft drinks, and the "authentic Indian food" category includes rice, curry, and other Indian foods. In both cases, the subcommodity includes foods with very different nutritional profiles, eliminating the possibility of linking one U-NDSR record to the entire category. About 2% of all purchases fell into homogeneous subcommodities but did not map due to the absence of a corresponding item in the U-NDSR database. Typically these were

locally prepared foods like macaroni salad or a meat and cheese tray, where ingredients vary depending on the method of preparation.

The last barrier to mapping nutritional information to the subcommodity categories was an occasional lack of understandable descriptions in text fields within the supermarket dataset. For example, the description "GIVE 2 BT PTRTC BRWN TRY" could not be matched to a U-NDSR food item since the meaning is unclear. This barrier was extremely uncommon, accounting for less than 1% of the unmapped categories. A subanalysis of heterogeneous subcommodities, which accounted for 21% of the unmapped subcommodities and nearly 30% of the entire data set, was conducted to improve our strategy for nutritional linking. The distribution of unique food items in this unmapped subpopulation showed a rapidly decaying distribution, suggesting that a small number of items constituted a large portion of the heterogeneous subcommodity population (Figure 4.4). The 10 most frequently purchased items in the distribution accounted for 9.6% of all 580,000 purchases from this large unmapped group. The 100 and 1,000 most frequently purchased items accounted for 25% and 63% of purchases, respectively. It is apparent that the majority of the purchases in this unmapped population represented a relatively small number of items.

The mapping of the U-NDSR food groups to the sales data was completed for 100% of the subcommodities in the supermarket data set. The food group was directly imputed from the U-NDSR food item code for the 70% of subcommodities that were successfully mapped to a U-NDSR food item. The remaining 30% of the subcommodities were manually assigned to a food group.

4.4 Discussion

We found that by using the supermarket's hierarchical system, a majority of sales data could be mapped to nutrient and food group values. Food groups were mapped to 100% of the sales data and nearly 70% of the entire dataset was linked to detailed nutritional content. To completely utilize sales data as a nutritional assessment tool, an analysis of the barriers encountered in mapping nutritional content and development of a strategy for increased linkage is needed.

The largest barrier to mapping the supermarket sales data to U-NDSR nutrient information was the presence of heterogeneous subcommodities containing nutritionally diverse food items. A potential solution is matching individually each food item in the unmapped categories to a U-NDSR entry. Manually mapping individual items in the heterogeneous categories would be straightforward, as it would be an identical process as the mapping of the subcommodities. Due to the distribution of the large unmapped population, shown in Figure 4.5, the most commonly purchased items account for a large percentage of the total unmapped items. With roughly the same effort of the original manual mapping that achieved 70% food item coverage, mapping the most frequently purchased 1,200 unmapped items to U-NDSR would extend coverage to 90% of all food items. This method would extend coverage considerably, leaving out only items that are purchased infrequently.

The second most common barrier, accounting for only 2% of the total purchases, was due to the lack of a representative item in the U-NDSR database. While some foods might not be present in the U-NDSR database, there are many other sources of nutritional information, such as food composition tables and online databases, where the information

for a representative food item for the subcommodities might be found. By using an alternate data source for these items, we can increase the linkage by 2%.

The supermarket's database structure was found to be adequate for mapping food groups to the sales data. The sales data were able to be mapped to food groups much more readily than nutrient information because the USDA food groups are very broad categories that fit the supermarket's definition of subcommodities well. For example, subcommodities such as "Authentic Indian Food," "Frozen Meal Dinners," and "Soft Drinks" contain foods too varied nutritionally to assign one nutritional profile to the entire category. However, the USDA food groups are broad enough that these subcommodities can fit into "Ethnic Foods," "Meals, Entrees, and Sidedishes," and "Beverages," respectfully. Similarly, the subcommodities with food items not found within the U-NDSR were easily categorized into a food group.

One limitation of this study is that the mapping we present is specific to the structure of the supermarket database that was used. Each supermarket chain will have a different structure and the process will need to be replicated for application to another database. In addition, due to the nature of the method used for mapping, the linkage was not 100% complete. Another limitation is that the U-NDSR is a proxy for nutritional content of the sales data, not the exact values for each product.

Despite the limitations, supermarket sales data offer several advantages that make it a unique data source. Sales data provide a benefit that is uncommon among other nutritional assessment methods: a steady stream of data that is very inexpensive to capture. Many assessment methods provide a one-time nutritional snapshot of an individual, household, or population (21). However, sales data are constantly being

collected, offering a rare view of the flow of foods over a period of time. Also, participant and researcher burden must be considered when choosing an assessment method (17). Sales data provide an indirect measure with virtually no participant burden and low researcher burden. In addition, as a consequence of using an indirect measure, researchers are rid of self-reporting biases that trouble other methods.

The ability to map nutritional content to sales data with reasonable effort allows researchers new opportunities. In a recent review, French et al. identified only two articles that utilized UPC scanners as the methodology to assess household purchasing behavior (17, 31, 55). One of the studies found that working with sales data at the UPC level is very difficult because new codes are constantly being created (55). Comprehensive databases are difficult to find and must be updated on a continual basis to provide the coverage needed. The investigators concluded that restrictions from using UPCs along with the complexity of the supermarket system result in extremely time consuming methods that are easy to underestimate. In additional studies that used purchasing data, researchers often selected subsets (e.g., fruits, vegetables, dairy, or high-fat foods) due to the immensity of data and the organizational challenges involved (17). In contrast to these previous studies, utilizing the internal categorization system within the supermarket database use has the potential to be an easily maintained, cost effective method of linking sales and nutritional data. The linkage need only be completed once, and through the use of less-granular categories instead of individual foods, the daily or weekly item changes within a store that would be necessary to track in a UPC database are avoided.

Successfully mapping nutritional information to sales data allows it to be used for many purposes. As sales data are routinely collected, public health surveillance can benefit from sales data as a means to determine regional or ethnic differences in eating behaviors (39, 43, 56, 57). Sales data can also be used in interventions to determine the effect of policy changes or healthy eating campaigns on purchasing patterns. Tracking purchasing patterns before and after the implementation of an intervention or policy change has the potential to be an indirect, low-cost tool (55, 58). In addition, monitoring household or individual purchases via sales data provides a potential means for dietary analysis and a feedback mechanism to inform and facilitate behavior change. The act of purchasing items is a conscious nutritional choice and thereby serves as a nutritional assessment method for individuals or households. Overall, the linking of supermarket sales data to the U-NDSR database shows potential as a useful tool with many applications in the public health, nutrition, and medical fields.

4.5 Conclusions

We showed that mapping nutrient information to sales data using a supermarket's built-in hierarchical structure is feasible. Successful mapping allows sales data to be utilized for many purposes including public health surveillance, intervention assessment, and household or individual food purchase evaluation.

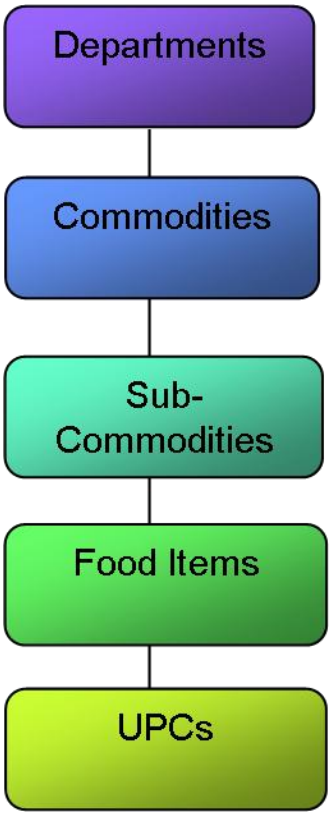

Figure 4.1 Structure of the supermarket database hierarchy

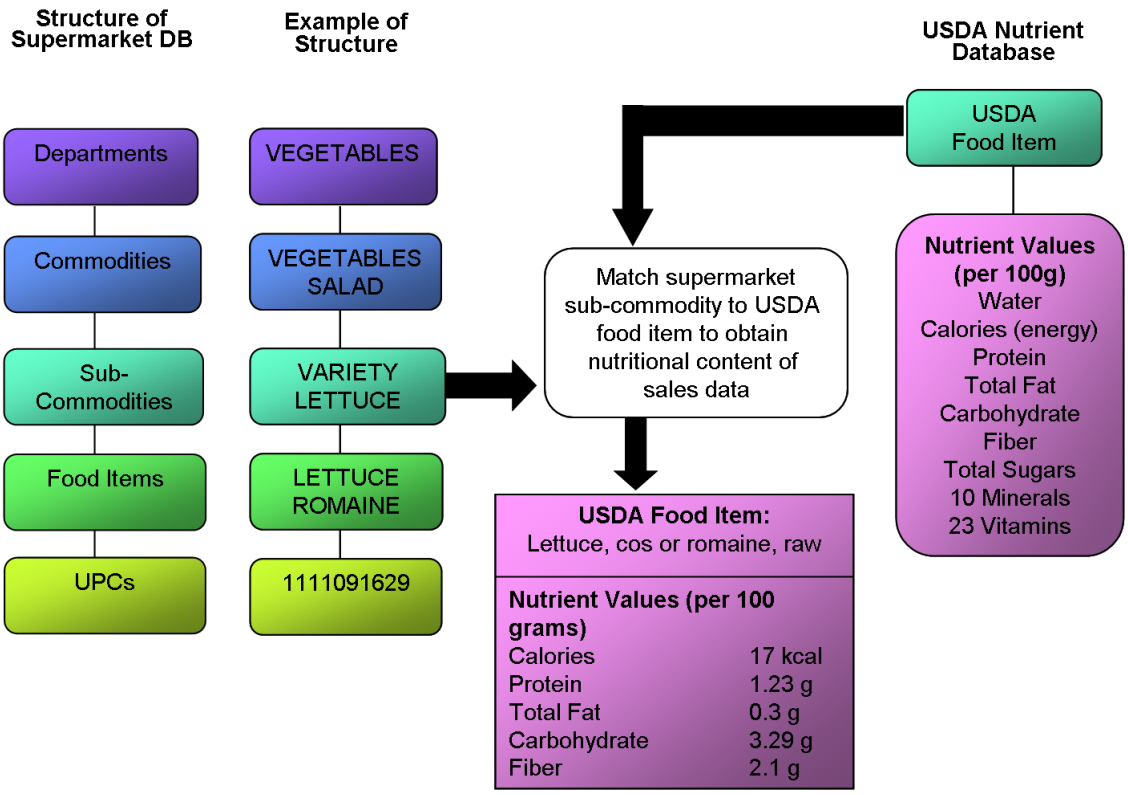

Figure 4.2 Diagram of mapping nutritional content to supermarket sales data.

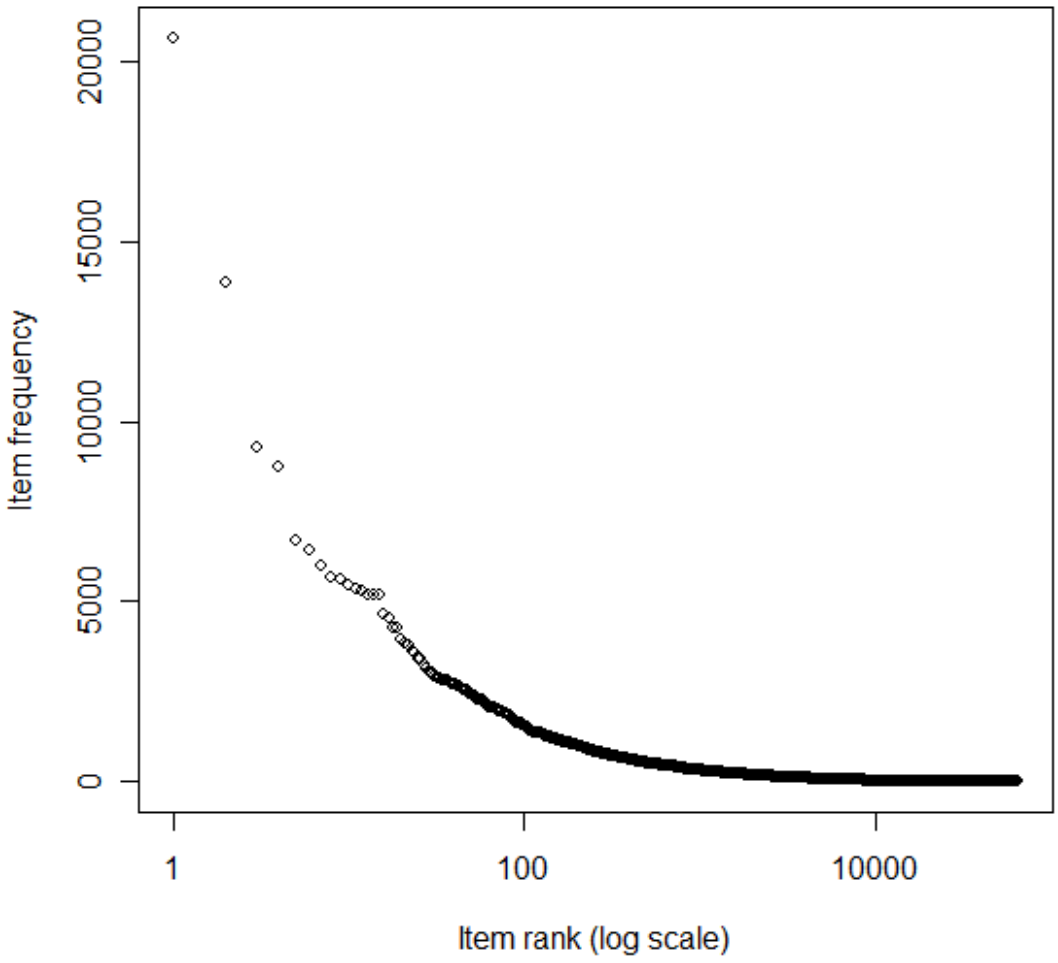

Figure 4.3 Distribution of items purchased in total data set

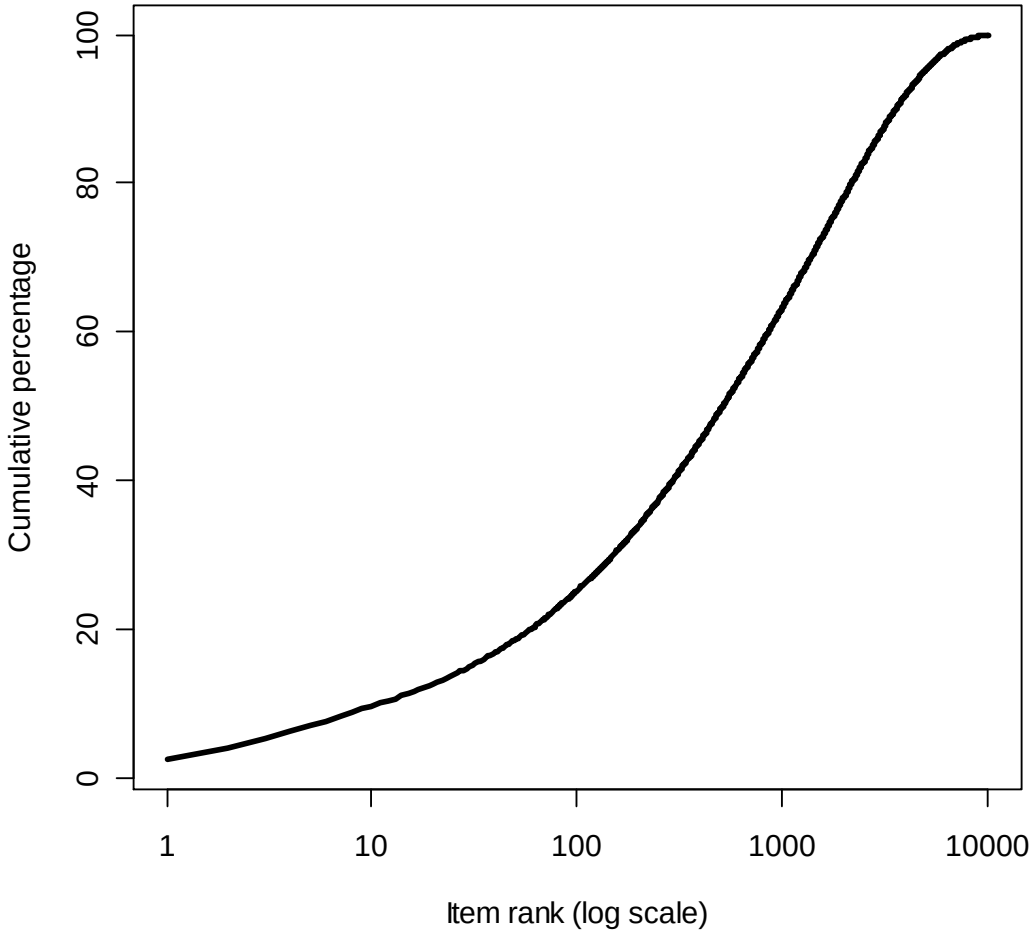

Figure 4.4 Distribution of unique food items from the large unmapped population.

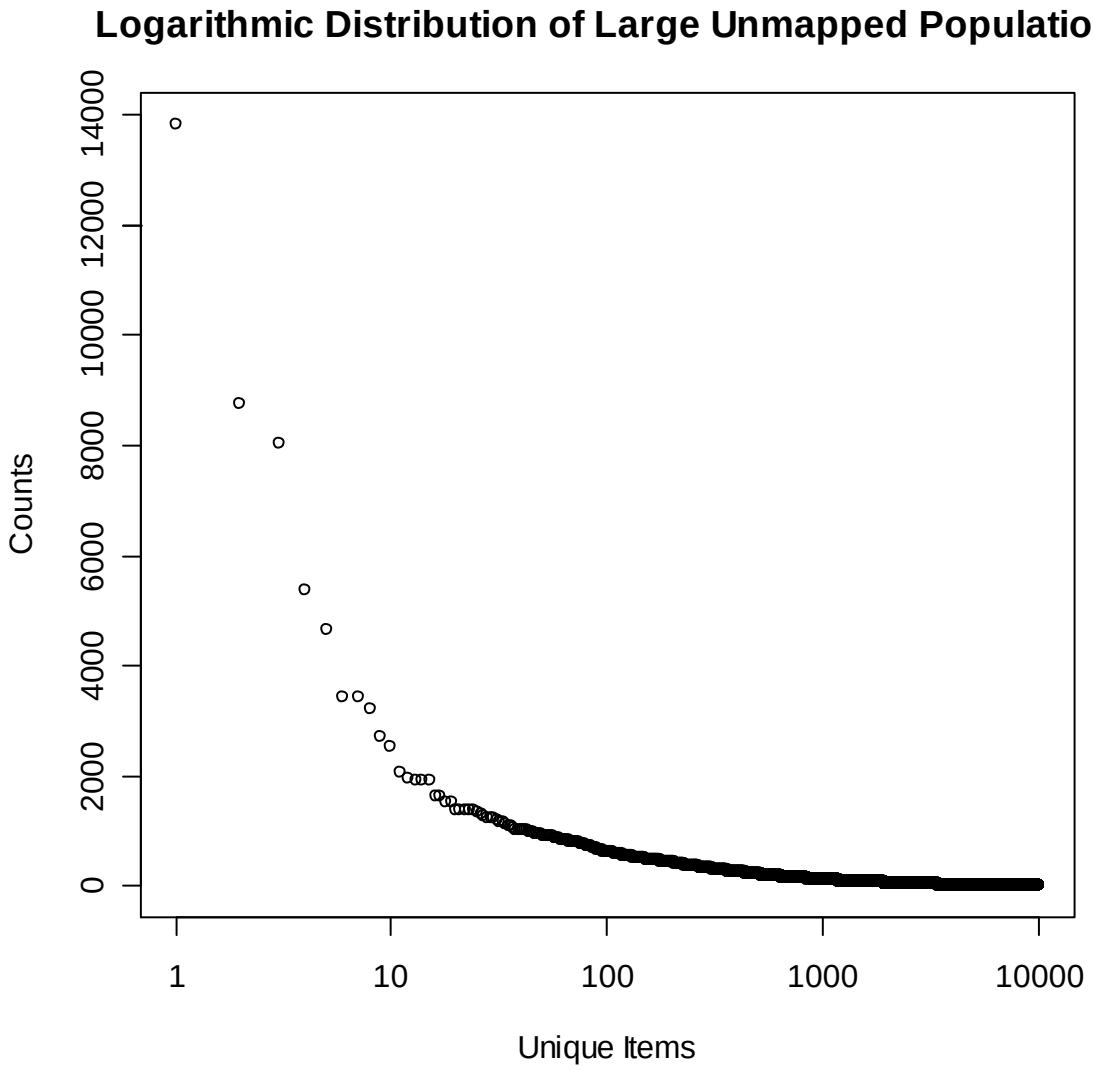

Figure 4.5 Distribution of large unmapped population.

Table 4.1 List of food groups used in USDA Nutrient Database

Dairy and Egg Products	Beef Products
Spices and Herbs	Beverages
Baby Foods	Finfish and Shellfish Products
Fats and Oils	Legumes and Legume Products
Poultry Products	Lamb, Veal, and Game Products
Soups, Sauces, and Gravies	Baked Products
Sausages and Luncheon Meats	Sweets
Breakfast Cereals	Cereal Grains and Pasta
Fruits and Fruit Juices	Fast Foods
Pork Products	Meals, Entrees, and Sidedishes
Vegetables and Vegetable Products	Snacks
Nut and Seed Products	Ethnic Foods

Table 4.2 Results of mapping nutrient data to sales data

	Subcommodity Level Results	**Unique Food Item Results**	**Entire Data Set Results**
Total number of food-related items	1,252 food-related subcommodities	26,854 unique food items	2,009,533 total food items
Number (percent) successfully mapped	884 (70.6%)	15,385 (57.3%)	1,387,864 (69.1%)
Number (percent) not mapped	368 (29.4%)	11,469 (42.7%)	621,669 (30.9%)
Not mapped due to nutritionally diverse subcommodity	263 (21%)	10,233 (38.1%)	580,768 (28.9%)
Not mapped due to subcommodity not being in USDA Database	103 (8.2%)	1,232 (4.6%)	40,892 (2%)
Not mapped due to inadequate text description	2 (0.2%)	4 (< 0.1%)	9 (< 0.1%)

CHAPTER 5

CORRELATION STUDIES

5.1 Objective

The objective of the work described in Chapter 5 is to use the correlation coefficient to explore the relationship between supermarket sales data and two nutritional measures: dietary intake as measured by a food frequency questionnaire and household food availability as measured by a household food inventory.

5.2 Methods

After accomplishing the first aim presented in Chapter 2 by recruiting a supermarket chain, successfully collecting sales data, food frequency questionnaires, and household inventories for the 50 households enrolled in the study, and mapping nutrient and food group data to the supermarket sales data, we were prepared to carry out the second aim; to study the relationship of sales data to traditional measures of a household. This aim was accomplished by employing correlation statistics to explore the relationship between sales data and household inventory and the relationship between sales data and food frequency questionnaire results.

Correlation was the first method of analysis chosen because it is a commonly used statistic to measure the relationship between two variables. Utilizing sales data as a method of nutritional assessment is still a largely underexplored area; therefore it is logical to begin the path of analysis with a simple, straightforward option, such as the correlation coefficient.

For three main reasons, we chose to focus solely on food group estimates for sales and household data throughout our analysis. The three reasons are: 1) the successful mapping of all food group data from the U-NSDR allows for complete food group coverage for the sales data, while the low mapping rate of nutrient data from the U-

NDSR would be a major limitation and would leave many purchased items without nutrient information; 2) the lack of unit size information for the sales data does not provide a means for calculating weight, volume, or number of servings per container for purchased items, which are necessary for determining the detailed nutritional content of each item purchased; and 3) studies are increasingly pointing to foods and food groups instead of individual nutrients as a research tool since the relationships between nutrients and chronic diseases are hard to solidify and individuals chose foods instead of nutrients when deciding what to eat (59, 60).

5.2.1 Sales Data in Food Group Form from Mapping

Sales data organized by food groups for the 50 households were obtained through the mapping with the U-NDSR described in Chapter 4. Specifically, we obtained sales data from the Large Intermountain Grocer comprising twelve months prior to the household visit for each household (n=50) in the study. As recent literature using sales data as a means of nutritional assessment does not describe an ideal time frame for aggregating and analyzing household purchases from a supermarket, the sales data were evaluated at 4 week cumulative intervals starting at the date of the household visit and running 52 weeks prior to the household visit to determine whether there was a single best time frame for correlation.

The sales data were analyzed according to percentages of items purchased in each food group. Raw numbers of items purchased within food groups would not suffice, as the total number of items purchased by each household varied. To negate the effect that the total number of items purchased would have on the analysis, the data were normalized by calculating the percentage of items purchased within each food group.

Normalization allows for all households to be analyzed on the same scale and gives an idea of the general makeup of household purchases by food groups.

5.2.2 Household Inventory Correlation Analysis

The data obtained from the household inventory as described in Chapter 3 included UPCs from food items with barcodes and a handwritten line item report of food items without barcodes. To compare percentages of foods in each food group within the household to food group percentages purchased, each food item included in the household inventory needed the correct food group assigned to it. A UPC database created and maintained by TrainingPeaks was used for this purpose.

The TrainingPeaks proprietary UPC database was used to obtain information on items with a UPC. The TrainingPeaks database consists of over 50,000 food items linked to nutritional content through the food items' UPC (source). Included in the TrainingPeaks database are detailed nutritional information such as calories, carbohydrates, protein, and fat, along with additional information such as servings per container and food group information. For the purposes of this analysis, only the food group information for food items was used. While the TrainingPeaks database includes detailed nutritional content for many UPCs, not all UPCs from food items found during the household inventories were present within the database. Items not in the TrainingPeaks database were manually entered into the database using photographs of the Nutrition Facts food label taken during the household visit. Nutrient and food group data for items without a UPC were mapped to a food item and the corresponding food group within the U-NDSR.

As described above for sales data, food group percentages for each household were calculated using all of the food items recorded during the household inventory. The correlation coefficients between the household inventory food group percentages as a single measurement and each of the four week cumulative time periods of the sales data were calculated to determine the trend of the correlation coefficient over the course of the preceding 52 weeks and if there was an ideal time frame for measuring retrospective sales data.

5.2.3 Food Frequency Questionnaire Correlation Analysis

The FFQ measuring dietary intake given to the mother of the household during the household visit was used for the correlation analysis. The self-reported raw values from the questionnaire were analyzed using DietCalc, a FFQ analysis software developed by the National Cancer Institute (DietCalc reference). DietCalc was created to be fully customizable to the FFQ being used, allowing a researcher to adjust the analysis parameters to match each question within the FFQ. DietCalc was used for this portion of the analysis for two reasons: it provides a unique ability to modify the software to fit the FFQ, and it gives the FFQ results in both nutrient and food group format, unlike many other software products or methods of analysis that provide nutrient only results.

As with the household inventory correlation analysis, the FFQ food groups were compared to the supermarket sales purchases for the year preceding the household visit using the correlation coefficient statistic. The food group categories in the output of DietCalc are not exactly the same as the food group categories from the U-NDSR, which was used for the sales data. Therefore, the correlation coefficient was calculated only for the food groups that do correspond with one another. Food categories from the DietCalc

analysis of the FFQ include dairy, discretionary fat, grains, fruit, meat, nuts and seeds, teaspoons of added sugar, and vegetables. The dairy, discretionary fat, fruit, nuts and seeds, and vegetable categories were compared to their direct food group counterpart within the U_NDSR method of categorization. The FFQ meat intakes for each household were compared to the combination of all meat related categories in the sales data (poultry, sausage and luncheon meats, pork, beef, finfish and shellfish, and lamb, veal, and game products), while the FFQ teaspoons of added sugar field was separately compared to both the Sweets and Beverages food groups in the sales data. The Sweets category was compared to the FFQ teaspoons of added sugar because the food group includes many food items with a high sugar content. The Beverages food group was compared to teaspoons of added sugar due to the fact that soft drinks are the leading source of added sugar in the U.S. diet (61).

Pearson's correlation coefficient was used to analyze the 52 week time frame to determine trends in the data and a possible best time period for comparison of sales and dietary intake data.

5.3 Results

Basic demographic information for the 50 households collected during the household visit is shown in Table 5.1. The comparison of household inventory food group percentages and supermarket sales data food group percentages are shown in Figure 5.1, which plots the correlation coefficient between the two variables vs. the week preceding the household visit. Overall, the graph shows that the correlation coefficient values are extremely variable in the first few weeks preceding the household visit and stabilize farther away from the visit and as data accumulate.

Over the course of the 52 week time frame, the correlation coefficients range from approximately r = -0.13 for the Finfish and Shellfish food group at week 52 up to r = 0.83 for the Baby Foods food group at week 20, with the vast majority of food groups falling in between r = 0.0 and r = 0.4. Along with the Baby Foods food group, there are two other food groups with an r value that rises above 0.5: Nut and Seed Products and Spices and Herbs. An additional five food groups cross the r = 0.4 threshold: Baked Products, Breakfast Cereals, Dairy, Snacks, and Soups, Sauces, and Gravies. Calculating the mean r values for each week averaged across all food groups, a minimum r value of 0.13 is found at week 4, while a maximum r value of 0.23 is found at week 32.

The graph showing the correlation coefficients between the FFQ food groups and supermarket sales data is shown in Figure 5.2. The correlation coefficients range from -0.17 at week four for the added sugars in the FFQ results compared with the sweets food group to 0.47 for the Meat food group. The correlation values display an early variability that leads to stabilization that is similar to the household inventory correlation graph. However, unlike the household inventory correlation analysis, the FFQ correlation does not have any r values over 0.5. Two comparisons yield r values greater than 0.4: the Meat food group, and sugar intake from the FFQ compared with percent of purchases from the Beverage sales data. When calculating the mean r values by week averaged across all food groups, there is a minimum at week 4 with an r value of 0.065, and a maximum r value of 0.20 at week 32.

5.4 Discussion

We found that there were several noteworthy results within the relationships between supermarket sales data and the household inventory and FFQ data. These results are discussed in more detail below.

5.4.1 Correlation Coefficients Analyzed by Average R Values

The average r value over the course of the 52 week time frame was at a maximum at week 32 and a minimum at week 4 for both the household inventory and FFQ portions of the study. While the average r value peaked at week 32 for both the household inventory and FFQ correlations, r values of 0.23 and 0.20 are not sufficient to say that there is a moderate correlation between the variables. In the comparison between supermarket sales data and the household inventory, the low averaged r value means that when analyzed by all food groups together, overall there is not a moderate correlation between household purchases from a supermarket and the items in a household. Although several food groups show moderate or high correlation, the average r value is not adequate enough to report any correlation.

There are several factors that might contribute to the r values not having a significant correlation for the household inventory data. First, items found in a household during one visit are not necessarily the exact proportions of what households are buying. For example, fruits are often consumed fresh and disappear from the house quickly because they have a short shelf life. Consequently, even though a large proportion of a household's purchases might be fruit, this food group might not represent as large of a proportion in the home as they do in the supermarket sales data. The data from the household visits reflect this idea. Out of the 50 households studied, 44 had a larger

percentage of fruits from sales data than from the household inventory; meaning the households purchased more fruits than what was represented in their house. Additionally, non-perishable items such as canned goods can stay in the house for an extended period of time because they have a long shelf life, and could result in being over-represented in the household data. Unfortunately, this relationship would be harder to tease out of the data, since many food group categories have canned items in them.

A second factor that might help account for the low r values is that some food items in the household are left in the cupboard or pantry because they are the food items that are *not* being consumed. During the household visits, it was not uncommon for a participant to comment that they did not even know that a specific food item was in their fridge, cupboard, or pantry. It is likely that many households purchase food items that are not regularly consumed, and thus sit in the house for months or even years. This observation can skew the household data to be more abundant for food groups that are purchased, but infrequently consumed.

Finally, households that regularly shop at supermarkets other than the Large Intermountain Grocer could negatively affect the correlation between sales data and household inventory. While the research team of this study worked with the Large Intermountain Grocer to recruit households that did the majority of their shopping at this specific grocer, it is unlikely that sales data from one supermarket chain precisely describes the households' shopping behaviors. Obtaining more information about food item purchases coming into the home from other sources might help describe the gap between household inventory and sales data.

In the comparison between supermarket sales data and the FFQ, the low average r value signifies that sales data analyzed by food groups for a household does not significantly correlate with a single measure of dietary intake from the mother of the same household.

There are several reasons that might help explain the low correlation coefficient. The FFQ is a one-time, self-reported measure of dietary intake from one member of the household while the supermarket sales data amasses purchases for every member of the household over an entire year. Obtaining additional dietary intake values for each member of the household would likely increase the correlation.

Another possible explanation for the low r value is that the results from the FFQ represent food items consumed from all possible sources, while the sales data only represent one source of dietary intake: a single supermarket. It is improbable that the proportions of food items or nutrients households purchase from each different source are exactly the same. Acquiring information about purchases from other supermarkets, convenience stores, fast food establishments, restaurants, and any other source of food would help account for dietary intake from other sources and possibly improve the correlation coefficient.

In addition, FFQs are known to have many inherent limitations, resulting in questionable accuracy. This study compared sales data against an imperfect measure: FFQs. Unfortunately, there is not a perfect measure of dietary intake to compare sales data against. However, more studies could be performed to compare sales data against other methods of measuring dietary intake, such as 24-hour food recalls and food diaries.

In looking back at the correlation analysis, another interesting finding is that both correlation studies have average r values that are a minimum at 4 weeks retrospective and peak at 32 weeks retrospective. While the peak r values for the household inventory and FFQ are only 0.23 and 0.20, respectively, the fact that they both have a maximum at 32 weeks shows that around 32 weeks of retrospective sales data is a suitable amount of data to include in a supermarket sales data analysis. As the time frame gets closer to 52 weeks, the average r value does decrease, although not drastically.

The decrease in the r value is likely due to the variability in purchasing behavior when including sales data from a time period that extends into the distant past. Although the HSFFQ used as the FFQ specifically asks respondents to recall their dietary intake over the past 12 months, results from FFQs are known to have questionable accuracy (16). While it might be reasonable for a respondent to summarize their dietary intake over the past month or two, reporting dietary intake averaged over the past 12 months is a more difficult task. Supermarket sales data are not stagnant measures; they are changing every time a household makes a trip to the supermarket. Therefore, it is likely that after a time period (around 32 weeks), supermarket sales data will not have the ability to correlate as well with one-time measures of dietary intake.

It is not surprising to find that the minimum r value for the comparison between sales data and both household inventory and FFQ occurs at four weeks retrospective. Figures 5.1 and 5.2 show that the correlation coefficient is much more variable in the weeks immediately preceding the household visit than in the remaining weeks. This is likely due to the fact that in the first few weeks preceding the household visit, there are not very many items purchased, and with a low item count, the variability between

observations can be great. The average number of food items purchased by the 50 households over the 4 weeks prior to the household visit is approximately 130, as compared to an average of 1,689 items purchased during the entire 52 weeks of the study. At four weeks prior, the sales data do not include a large number of data points and are less likely to be representative of average household purchases. At week 4, from the minimum r values of 0.13 and 0.065 for household inventory and FFQ, respectively, the small number of data points is not sufficient to adequately show a relationship between sales data and household inventory or FFQ.

5.4.2 Correlation Coefficients Analyzed by Food Groups

Separating out the analysis by food groups shows that several r values have a significant correlation. With a sample size of 50, an r value at or above 0.36 indicates significant correlation ($p < 0.01$). In the household inventory correlation study, the following food groups all have r values above the 0.361 threshold at least once during the 52 week time period: dairy, spices and herbs, baby foods, soups, sauces, and gravies, breakfast cereals, nut and seed products, baked products, meals, entrees, and sidedishes, and snacks. Out of the food groups with a significant r value, all except the meals, entrees, and sidedishes group have r values that reach above 0.4, with the baby food and spices and herbs food groups having r values above 0.5.

The Baby Food food group comparison between sales data and the household inventory yielded a higher r value than any other food group, with an r of 0.83 at 20 weeks of cumulative retrospective sales data. As seen in Figure 5.1, after the r value peaks at 0.83, it steadily decreases until it levels out around 0.65 between 44 and 52 weeks.

There are several unique features of baby foods that help account for the high correlation value of the Baby Food food group. One attribute of the Baby Food food group is that baby foods are generally either consistently purchased by a household or are not purchased at all. As baby foods have a specific purpose for a small percentage of the population (i.e., households with small infants), they are not an item purchased frequently by the majority of shoppers. Out of the 50 households participating in the study, 13 households did not purchase any food items within the Baby Food food group in the 52 weeks preceding the household visit. All 13 of those households also had no items within the Baby Food food group in the household inventory. This helps create a high correlation value, as there are many households whose sales data and household inventory match up exactly at zero for the Baby Food food group.

The peak at r = 0.83 with a subsequent decrease down to r = 0.65 can likely be explained by the fact that baby foods are only used in a household for a relatively short period of time. Baby formula is usually only used until an infant reaches his or her first birthday and strained or pureed baby foods and cereals are used from approximately 4 months of age until about 1 year or a little older. There are likely several households that would have purchases from the past year that fall into the Baby Food food group, but as their infant started eating solid foods, those households would have stopped purchasing and bringing baby food into their homes. Therefore, as the time between the date of the household visit and the cumulative sales data increases, the correlation between the one time household inventory and the cumulative sales data decreases, accounting for the large drop in r value as data gets closer to the 52 week mark. This is likely a unique

aspect of the Baby Food food group, as other food groups are regularly purchased by households month after month and year after year.

In the comparison between sales and FFQ data, three food category comparisons exceeded the significant r value of 0.361: total meats, nut and seed products, and added sugar intake compared with beverages purchased. The total meats food group comparison performed the best, reaching a correlation coefficient of 0.47 at 32 weeks' worth of retrospective sales data, while the nut and seed products food group is the only category that had a significant value for both the household inventory and FFQ portion of the study. The comparison between added sugar intake from the FFQ and beverages purchased from the sales data might seem like a surprise, however the number one source for added sugar in American's diet is soft drinks (61). The significant correlation between the two variables helps display that fact.

An action that would likely increase the correlation for both the FFQ and household inventory portions of this study is to improve the quality of supermarket sales data to eliminate as many limitations as possible. One limitation of the supermarket sales data we received is that they did not include information regarding container or food item size. Many popular products are available in several size options, which makes it difficult for researchers to accurately estimate the total nutrients in purchased foods. For example, when a customer purchases a box of Cheerios, the supermarket's database records that a box of Cheerios was purchased, and the price at which it was purchased, but the database doesn't record if the box is a 15 oz. or 20 oz. box. In addition, when that same customer purchases a carton of milk to use with the Cheerios, the type of milk (skim, 1%, 2%, whole, etc) is recorded, but the size of the carton (half gallon, gallon, etc) is not recorded.

Without product size information, researchers are left to estimate product size, use the percentages of total nutrients found in each food item to describe the general makeup of items, or use food group estimates as was used in this research.

Obtaining product sizes for produce items without a barcode are generally dealt with differently. When an item is weighed at the point of purchase (e.g., bananas), a weight is recorded and stored in the customer's transaction that can be used to determine the correct nutritional makeup of the purchased item. Information on the quantity of produce items that are paid for by count instead of weight (e.g., limes, celery stalks, and cantaloupes) is also recorded in the transaction. However, the quantity is not sufficient to determine the exact size of the food item. While many limes are approximately the same size, watermelons can differ greatly in size, and celery stalks might be available in several size options depending on the supermarket or season of the year when the purchase was made.

To garner more substantial data on products purchased, researchers have the possibility to urge supermarkets to gather the information in sales transactions. Although supermarkets gathering the data themselves might not be the likely solution, or the desirable solution, for supermarkets, it is a potential possibility and one worth exploring. Alternatively, information on product size could be gathered by researchers. This solution would likely not be comprehensive, as data that links UPC code to container size would need to be found for all food items in the supermarket database. Proprietary databases, such as the TrainingPeaks database, contain some information on product size that could be useful. Although, even in regularly maintained databases like TrainingPeaks, the data are not complete. If dealing with a small data set, it would be possible, although probably

not preferable, to fill in gaps by manually seeking out and obtaining the needed size information through images on the Internet or by physically searching for food items in a supermarket.

Another limitation of supermarket sales data that has been discussed briefly before is that sales data are not linked to exact nutritional information. While the majority of food items that have a UPC also include a Nutrition Facts Label that reports the exact quantities of nutrients in the food item, there is not a single, comprehensive database that links the two data sources together. This limitation leads to an estimation of nutrients based on other sources, such as the U-NDSR. An all-encompassing nationwide UPC database would be extremely useful to increase options for research and improve the quality sales data for nutritional analysis.

As examining the relationship between sales and nutritional data is a relatively new area of analysis, only one study with comparable results was found. Helen Eyles et al. compared the supermarket purchasing patterns of 49 primary household shoppers in New Zealand to 24-hour dietary recall results using the correlation coefficient as a method of analysis (62). While this study chose to analyze the data at the nutrient level instead of food group level, the range of r values obtained were similar with a maximum of 0.54 for percentage of energy from saturated fat and a minimum of 0.06 for sodium.

Although literature studying supermarket sales data is sparse, similar calculations have been performed between FFQs and other dietary assessment tools, and between repeated FFQs to check for reproducibility. While some of these studies reported maximum and average r values higher than our supermarket sales data correlation results, our results are in the same range of values and even very similar to a many of the

correlation coefficients reported in the literature (63-70). Many of the studies concluded that the correlation between assessment methods was good for foods or food groups consumed regularly and poor for those eaten infrequently. Further work with the supermarket sales data could be completed to determine if the food groups that ranked high or low possess unique qualities, such as frequency of consumption. Overall, the comparison between the supermarket sales data results and other reported results support the validation of using specific food group purchases from supermarket data as a surrogate for dietary intake.

One additional consideration that researchers in the nutrition field take is the effect of seasonality on dietary intake. As different foods are more readily available at different times of the year, the season in which measurements are taken can be biased. For example, certain fresh fruit and vegetable consumption increases in the warm summer months and fall harvest season when they are more readily available. This seasonality effect can be a big limitation of traditional nutritional assessment methods if the measure is not performed several times throughout the year. In contrast, supermarket sales data as a nutritional assessment method has the potential to collect data continually throughout the year. Continual data collection allows for yearly analysis that can account for seasonality, along with further analysis of seasonal purchasing habits.

A brief analysis of one years' worth of supermarket sales data was conducted to determine if any variation in purchases was visible over the year timeline. The percentage of food group purchases from all 50 households combined over the course of 12 months, from March 2007 to February 2008, is shown in Figure 5.3. The graph shows a high

month to month variation for many food groups and a possible seasonal variation for a few food groups.

The large month to month variation makes it difficult to visually examine a seasonal effect, however both the sweets and beverages food groups show a potential seasonal variation. The sweets food group shows a minimum from August to November, while the beverages food group shows a maximum peak from May to August. Further analysis on trends of sales data needs to be completed to determine what, if any, seasonal effect is evident in the food groups. Acquiring a full year of purchases from the supermarket for a larger population would allow for an improved seasonal effect analysis.

Weekly sale prices could play a large part in the month to month variation that is evident in many of the food groups. For example, the soups, sauces, and gravies food group spikes more than 10% in one month, accounting for 4.5% of the purchases in August to nearly 15% of all the household purchases in September. It is uncertain how much of that increase is due to regular variation; however it is likely that such a large spike is due to a sales promotion that occurred during the month of September.

Knowing that sales data has a large month to month variation is useful, as it shows that researchers cannot rely on simply 1 or 2 months of sales data for analysis. This conclusion corresponds with the results found earlier in the correlation study. The correlation study found that 32 weeks of data had the highest maximum r value, illustrating that researchers need to acquire and examine multiple months of supermarket sales data.

The findings of the analysis presented provide support for the use of supermarket sales data as a method for estimating the individual intake of certain food groups.

Supermarket sales data could be beneficial as a feedback mechanism provided by supermarkets for individuals or households who want to be more aware of food choices. In addition, dietitians and health care practitioners would benefit from receiving a summary of purchasing patterns to help patients on an individual basis. Supermarket databases could also potentially be used for the evaluation of interventions and public health programs and nationwide collection of sales data for studies such as NHANES or the National Children's Study.

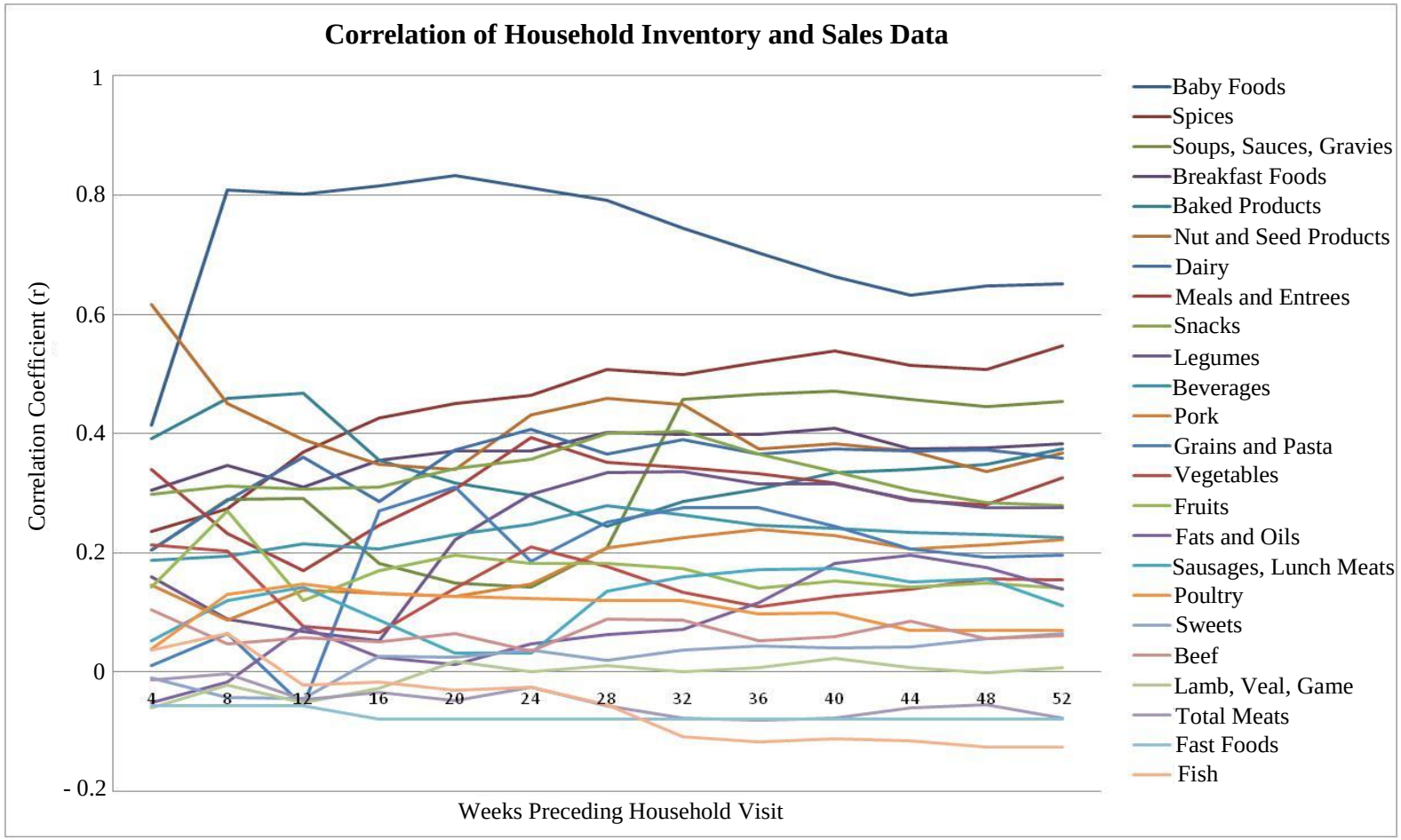

Figure 5.1 Correlation of household inventory and sales data. The food groups in the legend are listed from highest to lowest r value at 52 weeks.

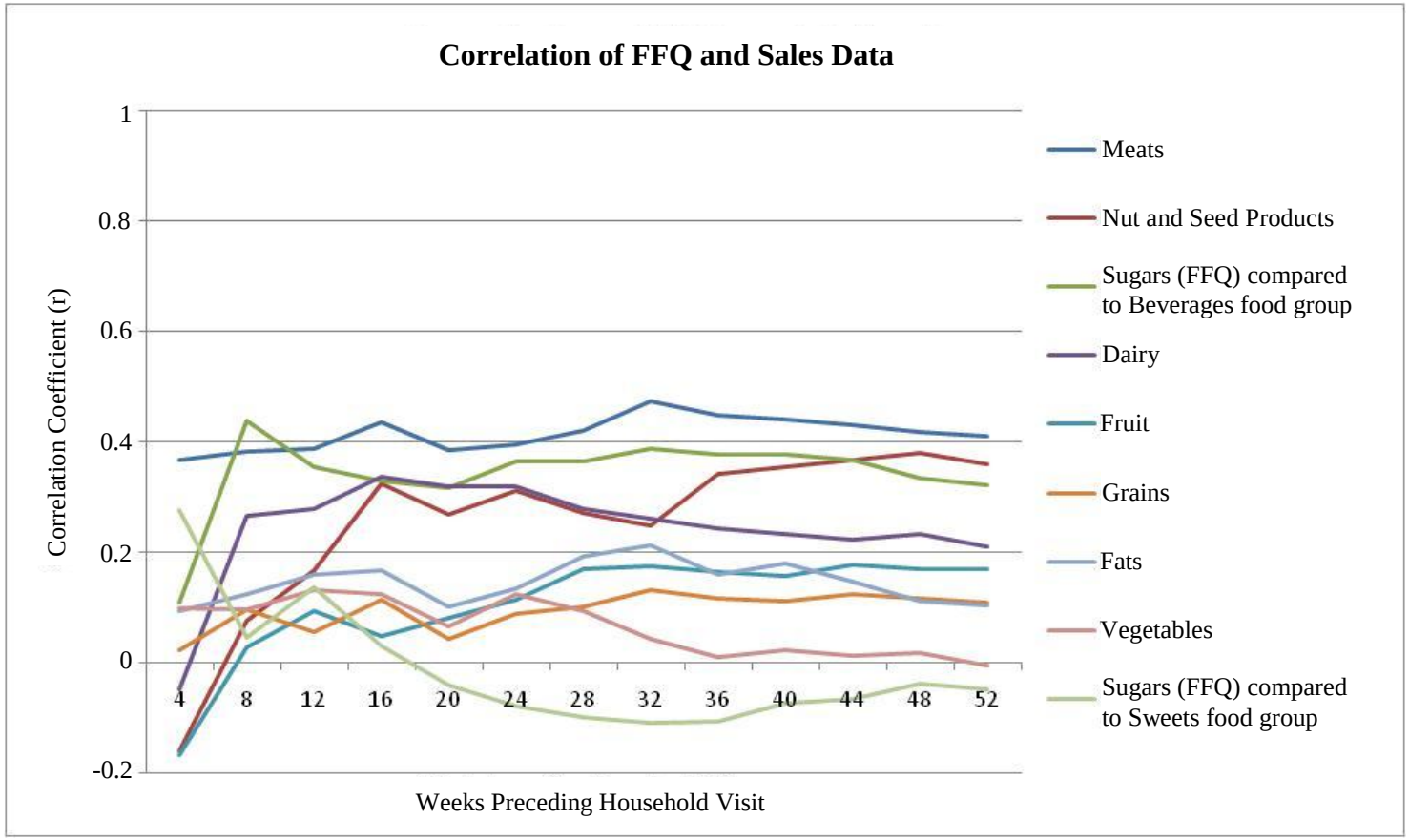

Figure 5.2 Correlation of FFQ and sales data. The food groups in the legend are listed from highest to lowest r value at 52 weeks

.

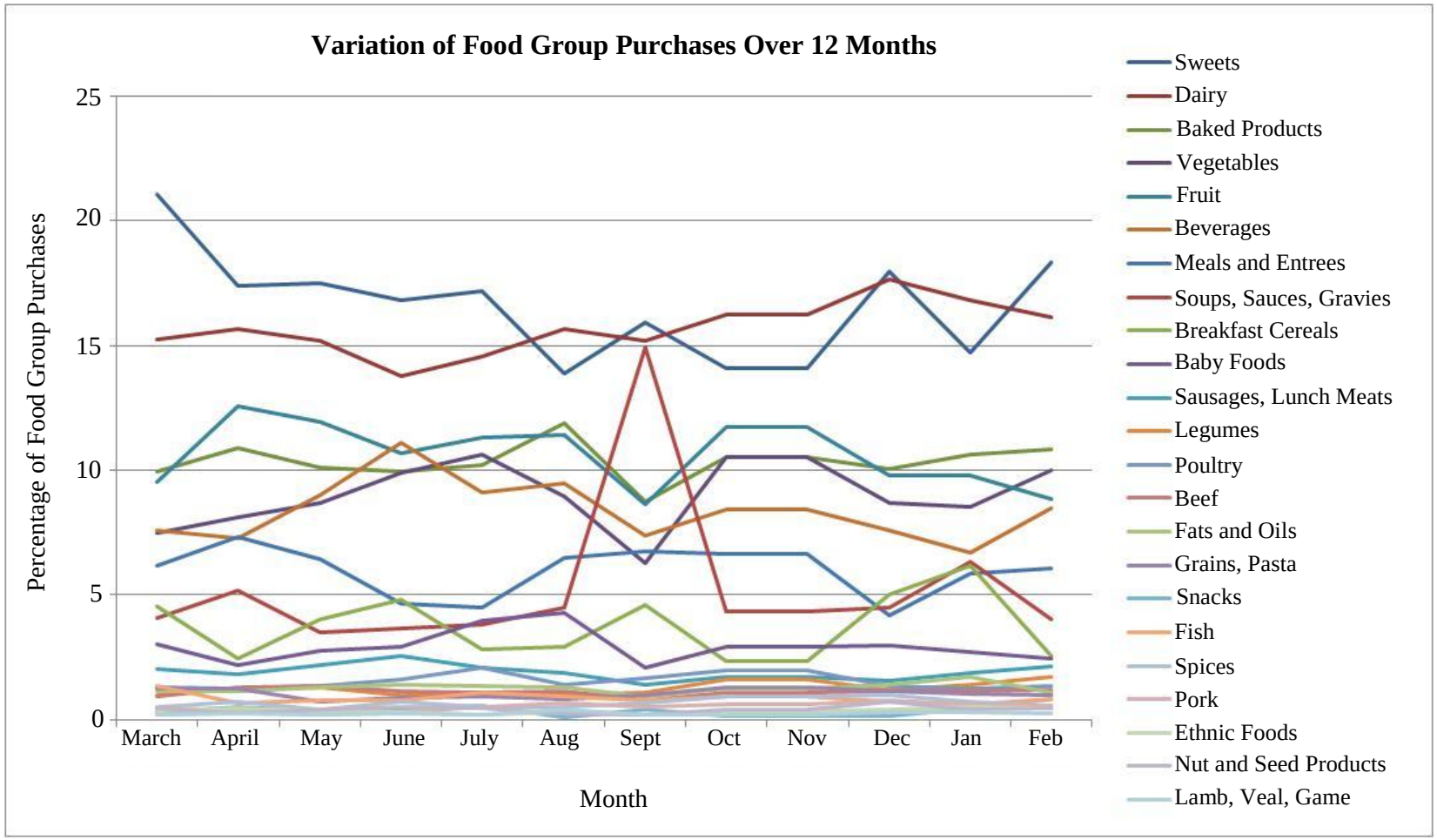

Figure 5.3 Variation of food group purchases for all 50 households combined over 52 weeks from March 2007 to February 2008. The food groups in the legend are listed from highest to lowest percent of food group purchases at 52 weeks

Table 5.1 Demographic Characteristics of the 50 Households

Characteristic	**N (%)**	**Mean (if applicable)**
Age of Mother		32.3
<26	6 (12%)	
26-30	14 (28%)	
31-35	13 (26%)	
>35	17 (34%)	
Age of Child participating		2.96
People residing in household		4.9
Adults		2.18
Children		2.72
Ethnicity of Mother		
Caucasian	48 (96%)	
Hispanic	2 (4%)	
Education Level		
< High school	1 (2%)	
High school	3 (6%)	
Some college	14 (28%)	
Associate degree	6 (12%)	
Bachelor degree	21 (42%)	
Graduate degree	5 (10%)	
Household Income		
<$15,000	2 (4%)	
$15,000-$29,999	5 (10%)	
$30,000-$44,999	7 (14%)	
$45,000-$59,999	11 (22%)	
>$60,000	25 (50%)	

CHAPTER 6

LOGISTIC REGRESSION MODELS

6.1 Objective

Multiple articles have described the association between individual dietary intake of food groups and health indicators such as body mass index (BMI), total energy intake (kcal or Calories consumed), and fat intake (71-75). These studies are important because efforts to reduce the prevalence of negative health outcomes associated with obesity should be more effective with an understanding of dietary patterns consistent with factors such as a high BMI or high fat intake. The literature describing dietary behavior generally uses conventional nutritional assessment methods such as 24-hour recalls and food frequency questionnaires to determine individual dietary intake. The aim of this chapter is to determine what relationships between dietary intake and BMI, energy intake, fat intake, and saturated fat intake can be found when using supermarket sales data as the measure of dietary intake in place of the traditional FFQ.

6.2 Methods

Body Mass Index (BMI) is a measure frequently used as an indicator of body fatness. BMI is calculated using solely an individual's height and weight and is consequentially an inexpensive, low-burden method used in many studies for determining weight categories that may lead to health problems (76).

Using the standard weight status categories from the CDC (76) in Table 6.1, the mother of each household was categorized into the correct weight status based on the self-reported height and weight given to the research team during the household visit. From the original four weight categories, the BMIs were combined into two categories for analysis purposes: Normal (BMI < 25.0) and Overweight (BMI ≥ 25.0).

Total energy intake is a measure frequently used to describe the total daily calories consumed by an individual. Energy intake is important because it directly affects an individual's weight. As energy intake increases above energy expenditure, an individual's weight will increase, and vice versa. Calculating daily energy intake compared to recommended intake is an important part of nutritional studies, as researchers can gain insight into general dietary behavior. Individual total energy intake is often measured through 24-hour food recalls, FFQs, and food diaries.

In this analysis, the estimates of daily calories consumed were calculated from the Mothers' FFQ results. Total energy intakes were then compared to the total energy expenditure (TEE) recommendations for each Mother of the household to determine the difference between energy intake and energy expenditure. TEE recommendations are based upon sex, age, weight, height, and regular physical activity (PA) level, using the following TEE equation for women (77):

$$TEE = 387 - 7.31 \times age + PA \times [(10.9 \times weight\ (kg) + 660.7 \times height\ (meters)]$$

The households were then categorized into two groups: the 25 households with the largest energy intake compared to recommended energy expenditure, and the 25 households with the smallest energy intake compared to recommended energy expenditure.

Fat intake is another frequently used nutritional measure. Individual fat intake is commonly assessed by a FFQ or 24-hour recall and is often reported as estimated grams of fat consumed per day or percent of daily calories consumed from fat. Fat intake is an

important measure because it has been consistently associated with negative health outcomes such as coronary heart disease and type 2 diabetes (74, 78, 79).

The Dietary Reference Intakes (DRIs) for macronutrients suggest that adult fat intake should fall between 20% and 35% of an individual's daily caloric intake (80). The DRIs were created by the Institute of Medicine's Food and Nutrition Board and are a list of recommendations of vitamins, minerals, and macronutrients for individual consumption organized by age and sex. The DRIs are used throughout the United States by health care professionals, researchers, and individuals to determine accurate estimates of dietary requirements. Using the DRI recommendations for fat intake, the 50 households were categorized into two groups based upon their fat intake as measured by the FFQ: meets DRI fat guidelines and exceeds DRI fat guidelines.

Saturated fat is a subset of fat intake and is an important nutritional measure because it has been linked to cardiovascular disease and multiple cancers, including colorectal, lung, and ovarian (81-85). Saturated fat guidelines from the DRIs state that an individual should keep their saturated fat intake "as low as possible while consuming a nutritionally adequate diet." (80) Based upon this recommendation, the 50 households were grouped into two categories based on the mothers' FFQ saturated fat results; the 25 households with the highest saturated fat intake and the 25 households with the lowest saturated fat intake.

Four separate logistic regression models were created to predict BMI status, energy intake, fat intake, and saturated fat intake from supermarket purchases. BMI status, energy, fat, and saturated fat were used as the dependent variables for the two models. The independent, or predictor variables, were the same in each model;

percentages of food groups purchased from the supermarket sales data. The food group data were gathered and organized as described earlier.

A stepwise logistic regression model was used to analyze the data. Stepwise regression models are unique in that they use an automatic procedure to choose the variables included in the final model. This is a useful and important feature for our analysis, as our sample size of 50 households limited the suggested maximum number of independent variables to five (86, 87). Accordingly, parameters within the regression models were adjusted to give five or fewer variables.

The data analysis for this portion of the paper was generated using SAS software, Version 9.2 of the SAS System for Windows. Copyright, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

6.3 Results

The distribution of the mothers' BMI for the 50 households is shown in the histogram in Figure 6.1. Combining the BMIs into two categories for analysis purposes yielded 27 households in the Normal weight category (BMI < 25.0) and 23 households in the Overweight category (BMI ≥ 25.0).

The results from the BMI stepwise logistic regression model are shown in Table 6.2. The stepwise model selected five food groups for the final model: spices, cereal, vegetables, beef, and meals (frozen entrees and side dishes). Out of the five variables in the model, only the vegetables food group showed a significant result with an odds ratio (OR) of 0.668 (95% CI: 0.470-0.949) and a p value of 0.0243. The spices,

cereal, beef and meals categories all contain non-significant p values and confidence intervals that include one.

The energy intake regression results are shown in Table 6.3. Fruit, fish, sweets, and meals were chosen for the final energy intake regression model. All four food groups included in the model had significant p values with ORs above one, indicating an increase in purchases of these food groups predicts a higher energy intake.

The distribution of fat intake by household is shown in Figure 6.2. Thirty one households were categorized into the "meets guidelines" group and 19 households in the "exceeds guideline" group.

The results from the fat intake stepwise logistic regression model are shown in Table 6.4. The model chose four food groups for the final model: luncheon meats, vegetables, fish, and sweets. All four of the chosen variables had significant results with p values below the 0.05 level. The fish food group was the only food group with an OR of less than one (OR: 0.117; 95% CI: 0.017-0.796), while the luncheon meats, vegetables, and sweets food groups had an OR above one.

The saturated fat results are shown in Table 6.5. The regression results contained five food groups for the final saturated fat model: beef products, beverages, legumes, luncheon meats, and sweets. Luncheon meats and beef products both had significant p values with ORs of 4.041 and 2.989, respectively. The legumes food group was the only significant food group with an OR below one (OR: 0.375; 95% CI: 0.143-0.983).

6.4 Discussion

The BMI study results revealed that 23 (46%) of the mothers of the 50 households were overweight or obese. The only food group that had a significant result was the

vegetable food group, with an OR of 0.668 (0.470-0.949). The OR, which is below one, indicates that a high percentage of purchases in the vegetables food group predict a normal BMI.

The knowledge that a high percentage of vegetables purchased by a household is predictive of a normal BMI is substantiated in articles that studied consumption of vegetables and BMI. In previous studies, vegetable consumption has been found to be inversely associated with obesity and other negative health outcomes like type 2 diabetes, cancer, and coronary heart disease (88-93). Many of these studies used traditional nutritional assessment methods such as 24-hour food recalls and FFQs to determine individual vegetable intake. The benefit of using supermarket sales data as a nutritional assessment method is that researchers can now potentially bypass the administration of lengthy, expensive questionnaires and collect food group purchases passively and automatically to predict BMI status.

Understanding trends in BMI over time is critical to monitoring the health of a population. Body mass index has been reported to be one of the most robust markers of diabetes, and studies have shown that "changes in BMI at the population level foreshadow changes in diabetes" (94-96). Having a mechanism to passively and automatically monitor trends in BMI through food group purchases could be extremely valuable in observing trends in BMI and subsequently diabetes.

The energy intake regression results yielded four food groups with significant results, which all had ORs above one, signifying an increase in the likelihood of inclusion into the higher daily caloric intake group. The results were inconsistent, with the sweets, meals, fruit, and fish food groups all predicting a higher energy intake. One might expect

the sweets and meals food groups to be associated with a high energy intake, as items within both the sweets and meals food groups consist of many calorically dense foods such as candy bars, ice cream, and frozen entrée meals. However, the fruit and fish food groups are generally associated with healthy eating behaviors and one would expect them to be associated with lower energy intake (97-99).

Looking at the main dietary sources of nutrients, one research groups listed the food items that constitute the top sources of energy among US adults (100). The table of top sources of energy is found in Table 6.6. Among the top 18 food items listed, two food items on the list, "potato chips/corn chips/popcorn" and "ice cream/sherbet/frozen yogurt," are items included within the supermarket sales data sweets food group. There are also several food items on the list of top sources of energy that would be included in supermarket purchases falling within the meals food group, such as beef, poultry, cheese, potatoes, pasta, and rice. The inclusion of these food items in the top sources of energy intake help substantiate the sweets and meals food groups as final predictors in the energy intake regression model. As expected, there are not any food items listed on the top sources of energy that would be included in the fruit or fish supermarket food groups.

There are several limitations in the energy regression model that likely contribute to the given results. One limitation is the use of FFQ data as the source of energy intake information for the 50 households. FFQs are completed by study participants and the results have been shown to have a bias toward healthy behavior (13-15). This self-reporting bias creates a problem as the households are grouped into either a high energy intake or low energy intake group for our analysis. As all traditional nutritional assessment techniques have innate weaknesses, it can be difficult to rely exclusively upon

the results from one study. Additional studies using supermarket sales data as a predictor for energy intake would be useful in determining which food groups have authentic relationships with energy intake.

In all of the analyses using FFQs, we also have to remember that the FFQ only represents the dietary intake of one member of the household: the mother. As supermarket sales data are a household measure of dietary behavior, there is likely to be some level of inconsistency between individual intake and the household supermarket purchases. Obtaining dietary intake information for all members of the household would be useful in helping to determine the level of concordance between individual dietary habits and household supermarket purchases.

Another limitation in our energy analysis is that we did not have the level of daily physical activity for the mothers of the households. The physical activity variable is needed to calculate individual total energy requirements and was therefore estimated in our analysis. We estimated all of the mothers to have the same physical activity level, which is likely not the case. Obtaining the correct physical activity level for each mother would result in more accurate energy estimates, and could also potentially change the classifications of the households into the high or low daily energy intake groups.

The fat intake study results showed that 19 (38%) of mothers from the 50 households have a daily fat intake greater than the recommended guidelines from the DRI tables. All four of the predictor variables chosen by the stepwise logistic regression model were significant. The luncheon meats, vegetables, and sweets food groups had an OR above one, indicating that they are predictive of a high fat intake. The fish food group

had an OR of less than one, indicating a predictive power of individual fat intake within the recommended guidelines.

The results of the fish food group are corroborated in many other studies showing that fish consumption has an inverse association with negative health outcomes, including coronary heart disease, myocardial infarctions, colorectal cancer, and breast cancer (71, 97-99). Given that fat intake has been shown to be positively associated with these negative health outcomes, it is logical that fish consumption is inversely related to both fat intake and certain negative health outcomes (74, 78, 79).

Out of the three food groups that had significant results with an OR above one, the luncheon meats and sweets food groups can be easily substantiated by multiple prior studies. Luncheon meats has been found to be one of the top ten sources of fat intake and high consumption of luncheon meats is frequently reported as being related to multiple types of cancer (75, 101, 102). The sweets food group, which includes food items such as candy bars, ice cream, and marshmallows, along with snack items such as potato chips, crackers, and popcorn, has also been shown to be a high component of dietary fat intake (72, 73, 100, 103).

As with the results from the BMI logistic regression model, discovering relationships between food groups purchased at a supermarket and fat intake can be useful. The luncheon meats and sweets results from the studies described above primarily used 24-hour food recalls as the method of gain individual dietary fat intake data. The knowledge that there are relationships between supermarket purchases and fat intake can inform future studies and new models for individual, household, and population level monitoring.

The final food group with a significant result in the fat intake regression model is the vegetable food group, with an OR of 1.469 (1.045-2.064). These results denote that a higher percentage of household purchases from the vegetable food group predict a higher individual fat intake for the mother of the household.

The unexpected association might be explained by how most Americans eat their vegetables: deep fried or topped with high fat dressings (104). In studying dietary behaviors that are associated with fat intake, several articles have reported that one of the top source of fat intake for US adults is salad dressings (75, 100). Additionally, an article examining US adolescent dietary behavior over three decades found that an increase in high fat potato consumption has resulted in an increase in adolescent vegetable consumption (105).

Explainable by high fat methods of vegetable preparation or not, the vegetable food group's significant results appear to contradict the results received from the BMI regression model, which found that the vegetable food group was predictive of a healthy demographic: a normal BMI. One would likely expect the vegetable food group to have an OR below one for both regression models, especially considering that fat intake has been correlated with BMI (74). When running a correlation analysis between BMI and fat intake for the 50 households in this study, the correlation coefficient was found to be -0.12, not the positive correlation found in literature.

Another limiting factor that might help explain the negative correlation is that both the BMI and fat intake measurements are composed from self-reported figures. The BMI variable is calculated from self-reported height and weight measurements, and the fat intake variable is determined from the FFQ completed by the participant. Self-

reported nutritional figures have been shown to have a bias toward the "healthy" behavior (13-15). Having a bias in both the BMI and fat intake is a limitation that might overwhelm the accuracy of the regression models.

The saturated fat regression model included two food groups that were predictive of high saturated fat intake: luncheon meats and beef. As stated earlier, luncheon meats has been found to be one of the top ten sources of fat intake and high consumption of luncheon meats is frequently reported as being related to multiple types of cancer (75, 101, 102). The inclusion of beef as a predictor of a diet high in saturated fat is similarly substantiated by many articles. Beef has been listed as the top source of dietary fat and the second source of saturated fat among US adults (100). Additionally, a high consumption of beef has been linked to negative health outcomes such as colorectal cancer and type 2 diabetes (106, 107).

The legumes food group was the only food group that resulted in a decreased likelihood of a high saturated fat intake. These results are in agreement with literature suggesting that a high consumption of legumes leads to a healthy diet (106).

A review of all four regression models shows that many food groups are included in multiple final models. For example, the beef, fish, luncheon meats, meals, sweets, and vegetables food groups are all included in more than one model, while the dairy and cereal food groups were purchased frequently by the households, but were not included in any models. This observation could indicate that certain food groups are more predictive of nutritional behavior or health outcomes than other food groups. Further research on this topic might point to important food groups to monitor compared to less important food groups.

In conclusion, we found that purchases of certain food groups can help predict BMI, energy, fat, and saturated fat intake. The findings in the logistic regression models give support to the potential of supermarket sales data as a source to monitor health factors such as BMI or nutritional intake. The ability to use supermarket sales data to predict population level trends of BMI or measures of nutritional intake has multiple advantages over traditional nutritional assessment methods. Traditional methods such as 24-hour dietary recalls and FFQs are time intensive for researchers and participants, expensive to administer and analyze, and can take months or years to collect and organize the results. Alternatively, supermarket sales purchases are an inexpensive data source that automatically collects data with little to no participant burden, while maintaining the potential to obtain and analyze data quickly.

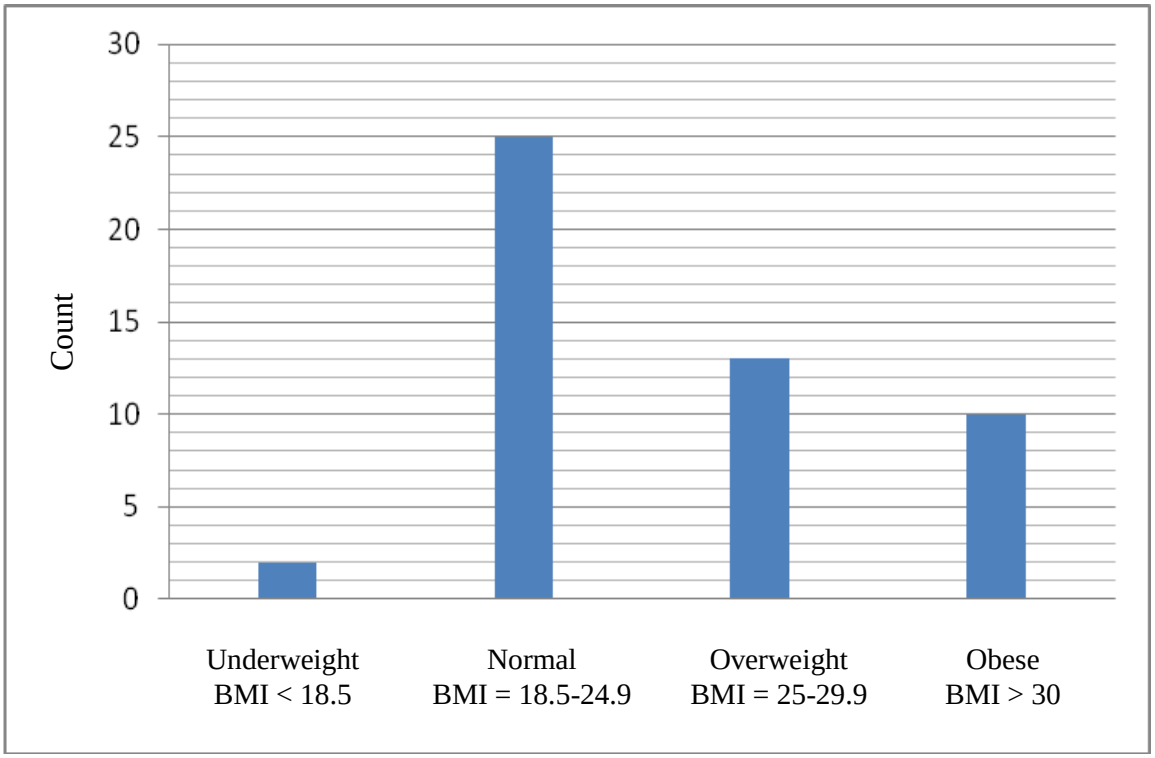

Figure 6.1 Histogram of BMI among mothers in 50 households.

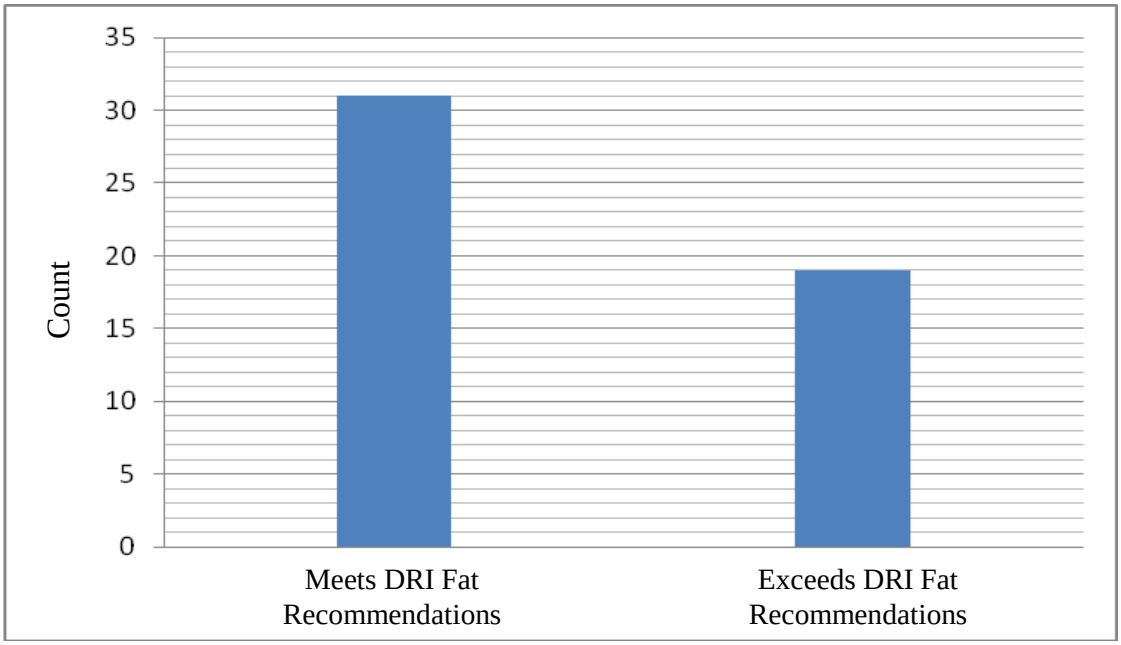

Figure 6.2 Histogram of fat recommendations among mothers in 50 households.

Table 6.1 Weight status categories based on BMI

BMI	Weight Status
Below 18.5	Underweight
18.5 – 24.9	Normal
25.0 – 29.9	Overweight
30.0 and above	Obese

Abbreviations: BMI, Body mass index (calculated as weight in kilograms divided by the square of height in meters)

Table 6.2 Logistic regression ORs for BMI

Food Group	OR (95% C.I.)	P Value
Spices	3.238 (0.982-10.671)	0.0535
Cereal	0.839 (0.613-1.150)	0.2748
Vegetables	0.668 (0.470-0.949)	0.0243*
Beef	2.439 (0.905-6.578)	0.0781
Meals (frozen entrees, side dishes, etc)	1.151 (0.889-1.490)	0.2851

Abbreviations: BMI, Body mass index (calculated as weight in kilograms divided by the square of height in meters); OR, odds ratio; CI, confidence interval
* Significant p value

Table 6.3 Logistic regression ORs for energy

Food Group	OR (95% C.I.)	P Value
Fruit	1.370 (1.048-1.791)	0.0214*
Fish	3.446 (1.019-11.655)	0.0466*
Sweets	1.140 (1.004-1.295)	0.0437*
Meals (frozen entrees, side dishes, etc)	1.444 (1.092-1.910)	0.0099*

Abbreviations: OR, odds ratio; CI, confidence interval
* Significant p value

Table 6.4 Logistic regression ORs for fat recommendation

Food Group	OR (95% C.I.)	P Value
Luncheon Meats	7.500 (2.177-25.836)	0.0014*
Vegetables	1.469 (1.045-2.064)	0.0268*
Fish	0.117 (0.017-0.796)	0.0283*
Sweets	1.289 (1.072-1.551)	0.0070*

Abbreviations: OR, odds ratio; CI, confidence interval
* Significant p value

Table 6.5 Logistic regression ORs for saturated fat

Food Group	OR (95% C.I.)	P Value
Luncheon Meats	4.041 (1.351-12.091)	0.0125*
Beef	2.989 (1.151-7.764)	0.0246*
Beverages	1.125 (0.969-1.306)	0.1227
Legumes	0.375 (0.143-0.983)	0.0461*
Sweets	1.110 (0.973-1.266)	0.1201

Abbreviations: OR, odds ratio; CI, confidence interval
* Significant p value

Table 6.6 Food sources of energy among US adults (100)

Ranking	Food Group
1	Yeast bread
2	Beef
3	Cakes/cookies/quick breads/doughnuts
4	Soft drinks/soda
5	Milk
6	Poultry
7	Cheese
8	Alcoholic beverages
9	Salad dressings/mayonnaise
10	Potatoes (white)
11	Sugars/syrups/jams
12	Pasta
13	Ready-to-eat cereal
14	Oils
15	Potato chips/corn chips/popcorn
16	Ice cream/sherbet/frozen yogurt
17	Rice/cooked grains
18	Margarine

CHAPTER 7

APPLICATIONS

7.1 Introduction

As supermarket sales data are applied to the nutrition and public health fields, a greater understanding of their capabilities are unveiled. Potential applications and areas of further exploration are described below.

7.2 Interventions

New research is pointing to supermarkets as an ideal location and data source for intervention studies (108). The work completed in this dissertation provides the building blocks for using supermarket sales data as a tool to create and evaluate nutritional interventions. One possible intervention that needs further study is the effect of pricing strategies on food and beverage purchasing habits. Ni Mhurchu et al. studied the effects that price discounts and education had on a population in New Zealand (50). While consumer education was not shown to have an effect on food choices, price discounts resulted in healthier food purchases after 6 months (50).

As sales data are not easily accessible to most researchers, several additional studies have reported the effect of pricing on food purchases using other data sources such as surveys and simulated supermarket laboratory environments (109-113). The analyses typically focused on modifying the prices of fruits and vegetables or high fat foods to influence customer purchases. While all articles reported promise in adjusting prices, it was also commonly mentioned that more research on interventions focusing on price discounts is needed. Using supermarket sales data would be an ideal data source for such interventions, as researchers would be getting sales data directly from the source.

Another area of further exploration in supermarket intervention studies is the effect of tailored education on consumers. Tailored education involves recording, gathering, organizing, and reporting individual sales data back to the consumer in a way that is easy to understand. The information can then be used by individuals and households to prompt healthy changes in diet. As changes in eating habits are often a result of individual responsibility, interventions focusing on helping individuals or households choose foods wisely might be the most successful. While a recent article describes the feasibility of collecting and disseminating personalized data to shoppers, there are little to no data on the effects that tailored education has on purchasing behavior (114).

7.3 Improve Individual Accountability

As sales data are studied more thoroughly, they have the potential to increase individual accountability. The potential to quantify sales purchases in an overall measure that is simple to understand, such as quality of diet, could result in dietary information being reported to clinicians as a means to encourage lifestyle discussions between patient and provider. For example, prediabetes identification and counseling in the primary care setting can be difficult (115). However, if clinicians have data that suggest a patient's diet is consistent with a lifestyle leading to diabetes, advice on modifying nutritional habits can prevent or delay the onset of diabetes (115). In the case of supermarket sales data, a simplified, easy to understand dietary record could be transferred to the clinician in the hopes of encouraging a dialogue about nutrition and dietary changes.

7.4 Nationwide Surveillance

Public health surveillance is another area that could be dramatically transformed by the application of informatics (116). Current nation-wide nutritional surveillance is limited to large scale studies such as NHANES that collect dietary intake measures through traditional methods like 24-hour dietary recalls, food diaries, and food frequency questionnaires. NHANES is a unique study that gathers a wide variety of data from participants; however the time frame from data collection to publication of results can be up to several years. For example, as of August 2011, the most recent NHANES data set available was NHANES 2007-2008 (117). Collecting nationwide supermarket sales data linked to nutritional information would be a huge boon to public health surveillance, as data collection and analysis could potentially be done in real-time.

Using the Real-time Outbreak and Disease Surveillance (RODS) system as a model for a nation-wide nutritional surveillance system could dramatically transform the field of nutrition. RODS collects de-identified data from healthcare visits in certain areas of the country (118). RODS is also home of the National Retail Data Monitor (NRDM), which collects data about over-the-counter (OTC) medication purchases from retail stores across the United States. The NRDM currently collects data on over 7,500 UPCs from OTC products used for self-treatment of infectious diseases (118). The data from OTC medication purchases and de-identified healthcare visits are combined, organized, and analyzed to produce relationships between medication sales and health-seeking behavior. Conclusions are then formed that can inform healthcare workers on the severity of seasonal diseases that can affect staffing along with medicine or immunization inventory.

For example, the RODS laboratory found that a large number of OTC electrolyte purchases preceded gastrointestinal and respiratory hospital visits by 2.4 weeks (119). Finding relationships between OTC purchases and health-seeking behavior are important because they can potentially be early indicators of outbreaks of diseases.

Similarly, a nation-wide surveillance system for food and beverage purchases could inform dieticians, clinicians, and public health workers on the state of well-being of the population in the United States, leading to greater preparation and hopefully, better health outcomes. Just as the RODS laboratory found that OTC electrolyte purchases precede GI and respiratory hospital visits, a surveillance system monitoring food and beverage purchases could find that purchases of certain foods or food groups precede health outcomes such as increased BMI, diabetes, or heart disease. An increase in BMI at the population level has been found to precede an increase in diabetes (94). Finding a sales data predictor for BMI, and subsequently diabetes, would allow professionals to better prepare for the corresponding consequences and formulate strategies to reverse the negative health outcomes.

The creation of RODS required a large amount of infrastructure to be built before the first hypotheses could be tested. Similarly, it is likely that a lot of work would need to go into building a large-scale food and beverage surveillance system before concrete results could be formed. The results presented in this dissertation display only a small sample of the potential relationships between supermarket sales data and health outcomes. Having a nationwide surveillance system would certainly provide novel and useful findings.

7.5 Importance to Informatics

Public health informatics is defined as the "systematic application of information and computer science and technology to public health practice, research, and learning" (116). Information science includes: analysis of structure, properties, and organization of information; information storage and retrieval; database architecture and design; and project management while computer science involves the theory and application of automatic data processing machines, pattern recognition, and artificial intelligence (116).

Many applications of information and computer science were used during the course of this dissertation. Analysis of structure, properties, and organization of information were heavily relied upon as the supermarket sales data were made available to us. There is very little information published on supermarket sales data, as they are generally not provided to researchers, and a detailed analysis of the structure and organization was vital to understanding the data, noting strengths and weaknesses, and realizing potential applications. Database design and information storage and retrieval were also important pieces of the research process. A database was created to efficiently organize and store the data elements for easy access, querying, and retrieval. Various aspects of computer science were also utilized to attempt automatic data processing in linking the USDA nutrient database to the supermarket sales data.

While computer and information science are key elements of informatics, public health informatics involves bringing together specialists from multiple disciplines to form new ways of solving public health problems (120). The work described here involved specialists and knowledge from many fields of study: nutrition, public health, clinical

medicine, and statistics, along with the traditional information science, and computer science disciplines.

A primary focus of public health informatics must be identifying diseases and predecessors of diseases in populations with simplicity and speed while disseminating new knowledge quickly and in ways that support current public health practice (120). This dissertation work focused on a novel method of nutritional assessment that has the capability to collect pertinent data quickly and automatically, allowing researchers to analyze and report trends in purchasing behavior that are related to health outcomes. As more research is done in the field of supermarket sales data, relationships between food purchases and health outcomes are likely to become more apparent and validated by further studies, allowing continual surveillance of supermarket purchases. These data and analyses have the potential to improve upon current methods of public health practice, as collecting, organizing, and analyzing nutritional data has traditionally been a laborious task not possible in real-time.

As some manual procedures were used in the linkage of USDA data to the supermarket sales data, one might wonder why we consider this analysis to be informatics. The work described follows the principle set forth by Yasnoff, et al that public health informatics is more than just automation, but "enables the redesign of systems using approaches that were previously impractical or not even contemplated" (116). The work of linking nutritional data to sales data lays the groundwork that begins to release the full potential of supermarket sales data as a nutrition and public health tool. By linking nutritional data to sales data, new projects, analyses, and systems that will potentially improve conventional public health practices are now practical. This project

truly allows the investigation of dietary intake, nutritional behaviors, and health outcomes associated with dietary intake, using a new method that has the potential to be used efficiently and cost-effectively on an individual, household, and population-level basis.

APPENDIX A

MATERNAL AND CHILD QUESTIONNAIRE

Maternal and Child Questionnaire

Please fill out this questionnaire, answering to the best of your ability for both you and your child. The child that you are answering for is the 1 to 5 year old participating in the study. Please keep this child in mind and answer for him/her whenever you read a question referring to "your child."

1. How many children under age 18 live in your house?

   _____

2. How many adults (anyone over age 18) live in your house?

   _____

3. What is your ethnicity? (Check one)
   _____ Native American
   _____ Pacific Islander
   _____ Asian
   _____ Black
   _____ Hispanic
   _____ White
   _____ Other: _____

4. What is the highest level of education YOU have completed? (Check one)
   _____ Partial high school or less
   _____ High school graduate / GED
   _____ Some college (including technical or vocational school), no degree
   _____ Associate degree
   _____ Bachelor's degree
   _____ Post-graduate degree
   _____ Other:_____

5. What is YOUR height? _____feet _____inches

6. What is YOUR weight? _____lbs.

7. What is YOUR date of birth (month/day/year)?
   _____

8. What is YOUR CHILD'S height? _____feet _____inches

9. What is YOUR CHILD'S weight? _____lbs.

10. What is YOUR CHILD'S date of birth (month/day/year)?
    _____

11. What was your total household income for 2007? Please include ALL ADULTS contributing to your total household income.
    _____ Under $15,000
    _____ $15,000 to $29,999
    _____ $30,000 to $44,999
    _____ $45,000 to $59,999
    _____ $60,000 or above

12. On average, how many times per week do YOU eat meals that were prepared by a restaurant? Please include eat-in restaurants, carry-out restaurants, fast-food establishments, and food delivered to your house.
    _____ Never
    _____ Less than once per week
    _____ Times per week (write in number of times per week)

13. On average, how many times per week does YOUR CHILD eat meals that were prepared by a restaurant? Please include eat-in restaurants, carry-out restaurants, fast-food establishments, and food delivered to your house.
    _____ Never
    _____ Less than once per week
    _____ Times per week (write in number of times per week)

14. Do YOU watch TV while you eat?
    _____ All of the time
    _____ Most of the time
    _____ Some of the time
    _____ Rarely
    _____ Never

15. Does YOUR CHILD watch TV while he/she eats?
    _____ All of the time
    _____ Most of the time
    _____ Some of the time
    _____ Rarely
    _____ Never

16. How much of your household food comes from hunting, farming, or gardening? Please do not include meat and produce purchased through a store.
    _____ All
    _____ Most
    _____ Some
    _____ Hardly any
    _____ None

17. Do you have food storage items in your house that are not part of your regularly used food items?
    _____ Yes
    _____ No

18. How often does your household shop at the following grocery stores per month? (write in the number by the name of each store)
    _____ Smith's Food and Drug
    _____ Albertson's
    _____ Dan's
    _____ Wal-mart
    _____ Sam's Club/Costco
    _____ Harmon's
    _____ Target
    _____ Other: _____
    _____ Other: _____

19. Are there certain types of foods or food items that you regularly buy at a store other than Smith's Food and Drug? If so, write the names of the food items by the name of the store that you buy it from. For example: if you frequently go to Sam's Club to buy milk, please write "milk" next to Sam's Club.

    Albertson's_____
    Dan's _____
    Wal-mart _____
    Sam's Club/Costco_____
    Harmon's _____
    Target_____
    Other Grocery Store: _____
    Other Grocery Store: _____

APPENDIX B

ADULT HARVARD SERVICE FOOD

FREQUENCY QUESTIONNAIRE

Name:_____
ID:_____
Date ____/____/____
DOB: ____/____/____

Check one:
☐ Pregnant
Due Date_____
☐ Breastfeeding
1st 6 months
☐ Breastfeeding
2nd 6 months
☐ Not Breastfeeding

Nutrition Questionnaire

What Have You Been Eating Lately?

During the past 4 weeks, how often did you eat a serving of each of the foods listed here?

Mark only one X for each food

Example:

	last 4 weeks		**each week**			**each day**			
Number of times	0	1–3	1	2–4	5–6	1	2–3	4–5	6+
Milk				**X**					
Hot chocolate	**X**								

	last 4 weeks		each week			each day			
Number of times	0	1-3	1	2-4	5-6	1	2-3	4-5	6+
Milk									
Hot Chocolate									
Cheese, plain or in sandwiches									
Yogurt									
Ice cream (cones, sandwiches, sundaes)									
Pudding									

What kind of milk do you usually drink? (Check one)

☐ whole ☐ 1% ☐ chocolate milk
☐ 2% ☐ skim ☐ other_____

	last 4 weeks		each week			each day			
Number of times	0	1-3	1	2-4	5-6	1	2-3	4-5	6+
Orange juice or grapefruit juice									
Other juice									
Fruit drinks (Kool-aid, lemonade, sportsdrink)									
Banana									
Peaches									
Fruit cocktail, mixed fruit									
Orange or grapefruit									
Apple or pear									
Applesauce									
Grapes									
Strawberries									
Melon									
Pineapple									
Raisins or prunes									

Mark only one X for each food.
How often did you eat a serving of these foods during the past 4 weeks?

	last 4 weeks		each week			each day			
Number of times	0	1-3	1	2-4	5-6	1	2-3	4-5	6+
Corn									
Peas									
Tomatoes, tomato sauce, salsa									
Peppers (green, red, or hot)									
Carrots									
Broccoli									
Green beans									
Spinach									
Greens (mustard, turnip, kale)									
Mixed vegetables									
Squash, orange or winter									
Zucchini, yellow squash									
French fries, fried potatoes, tatertots									
Potatoes (baked, boiled, or mashed)									
Sweet potatoes or yams									
Cabbage, coleslaw or cauliflower									
Lettuce salad									
Salad dressing									
Mayonnaise									

	last 4 weeks		each week			each day			
Number of times	0	1-3	1	2-4	5-6	1	2-3	4-5	6+
Chips									
Popcorn or pretzels									
Crackers									
Nuts									
Cookies or brownies									
Cake or cupcake									
Pie									
Jello									
Chocolate or candy bar									
Other candy (not chocolate)									
Coffee or tea									
Soda, soft drink, pop (not sugar free)									
Soda, soft drink, pop (sugar free)									
Beer, wine, wine cooler, mixed drink									

	last 4 weeks		each week			each day			
Number of times	0	1-3	1	2-4	5-6	1	2-3	4-5	6+
Beans (baked, chili, or other)									
Rice									
Spaghetti or other pasta									
Pizza									
Tacos, burritos									
Macaroni and cheese									
Hot dogs									
Sausage									
Hamburger (prepared any way)									
Canned tuna									
Fried fish, fish sticks									
Other fish									
Cold cuts (baloney, ham, salami)									
Fried chicken, chicken nuggets									
Other chicken or turkey									
Pork or ham									
Roast beef or steak									
Liver, organ meats									
Peanut butter									
Bread (slice) toast, roll, or pita									
Butter (not margarine)									
Margarine									

	last 4 weeks		each week			each day			
Number of times	0	1-3	1	2-4	5-6	1	2-3	4-5	6+
Vegetable soup									
Other soup									
Cornbread or tortilla									
Eggs									
Bacon									
Hot cereal, grits									
Cold cereal									
Donut									
Sweet roll or muffin									
Pancake, waffle, or French toast									
English muffin or bagel									
Biscuit									

1. What type of bread do you usually eat:

☐ white bread ☐ whole wheat or dark bread
☐ about half and half ☐ DON'T EAT BREAD

2. What type of margarine do you usually use:

☐ stick ☐ tub ☐ squeeze ☐ DON'T USE MARGARINE

Is this margarine:

☐ corn oil ☐ nonfat ☐ other

3. If you eat cold breakfast cereal, what type:

☐ high fiber (All Bran) ☐ other (e.g., Corn Flakes)

4. Do you take a multi-vitamin pill (Centrum, One-A-Day):

☐ no ☐ yes

If yes, how often:

☐ every day ☐ 4-6 times a week ☐ 1-3 times a week ☐ less than one time a week

5. Do you take a separate iron pill (not in the multi-vitamin pill above):

☐ no ☐ yes

6. Do you take a separate vitamin A supplement (not in the multi-vitamin pill above)"

☐ no ☐ yes

7. Do you take a separate calcium pill (not in the multi-vitamin pill above):

☐ no ☐ yes

8. Do you eat fried food at home:

☐ no ☐ yes

If yes, how often

☐ every day ☐ 4-6 times a week ☐ 1-3 times a week ☐ less than one time a week

If yes, what type of fat do you use to fry at home

☐ butter ☐ maragarine ☐ crisco ☐ corn oil ☐ canola oil
☐ olive oil ☐ other vegetable oil

9. Do you bake cookies, cake or pies at home:

☐ no ☐ yes

If yes, how often do you eat home-baked cookies, cake, or pies:

☐ every day ☐ 4-6 times a week ☐ 1-3 times a week ☐ less than one time a week

If yes, what type of fat do you use to bake at home:

☐ butter ☐ maragarine ☐ crisco ☐ corn oil ☐ canola oil
☐ olive oil ☐ other vegetable oil

APPENDIX C

CHILD HARVARD SERVICE FOOD

FREQUENCY QUESTIONNAIRE

Children's Nutrition Questionnaire

What Have You Been Eating Lately?

During the past 4 weeks, how often did you eat a serving of each of the foods listed here?

Mark only one X for each food

Name:_____

ID:_____

Date ___/___/___

DOB:___/___/___

_____

Age:_____

Respondent (please check):

☐ Mother

☐ Other _____

Example:

	last 4 weeks		each week			each day			
Number of times	0	1–3	1	2–4	5–6	1	2–3	4–5	6+
Milk				X					
Hot chocolate	X								

	last 4 weeks		each week			each day			
Number of times	0	1-3	1	2-4	5-6	1	2-3	4-5	6+
Milk									
Hot Chocolate									
Cheese, plain or in sandwiches									
Yogurt									
Ice cream (cones, sandwiches, sundaes)									
Pudding									

What kind of milk does your child usually drink? (Check one)

☐ whole ☐ 1% ☐ chocolate milk

☐ 2% ☐ skim ☐ other_____

	last 4 weeks		each week			each day			
Number of times	0	1-3	1	2-4	5-6	1	2-3	4-5	6+
Orange juice or grapefruit juice									
Other juice									
Fruit drinks (Kool-aid, lemonade, sportsdrink)									
Banana									
Peaches									
Fruit cocktail, mixed fruit									
Orange or grapefruit									
Apple or pear									
Applesauce									
Grapes									
Strawberries									
Melon									
Pineapple									
Raisins or prunes									

Mark only one X for each food.
How often did you eat a serving of these foods during the past 4 weeks?

	last 4 weeks		each week			each day			
Number of times	0	1-3	1	2-4	5-6	1	2-3	4-5	6+
Corn									
Peas									
Tomatoes, tomato sauce, salsa									
Peppers (green, red, or hot)									
Carrots									
Broccoli									
Green beans									
Spinach									
Greens (mustard, turnip, kale)									
Mixed vegetables									
Squash, orange or winter									
Zucchini, yellow squash									
French fries, fried potatoes, tater tots									
Potatoes (baked, boiled, or mashed)									
Sweet potatoes or yams									
Cabbage, coleslaw or cauliflower									
Lettuce salad									
Salad dressing									
Mayonnaise									

	last 4 weeks		each week			each day			
Number of times	0	1-3	1	2-4	5-6	1	2-3	4-5	6+
Chips									
Popcorn or pretzels									
Crackers									
Nuts									
Cookies or brownies									
Cake or cupcake									
Pie									
Jello									
Chocolate or candy bar									
Other candy (not chocolate)									
Coffee or tea									
Soda, soft drink, pop (not sugar free)									
Soda, soft drink, pop (sugar free)									

	last 4 weeks		each week			each day			
Number of times	0	1-3	1	2-4	5-6	1	2-3	4-5	6+
Beans (baked, chili, or other)									
Rice									
Spaghetti or other pasta									
Pizza									
Tacos, burritos									
Macaroni and cheese									
Hot dogs									
Sausage									
Hamburger (prepared any way)									
Canned tuna									
Fried fish, fish sticks									
Other fish									
Cold cuts (baloney, ham, salami)									
Fried chicken, chicken nuggets									
Other chicken or turkey									
Pork or ham									
Roast beef or steak									
Liver, organ meats									
Peanut butter									
Bread (slice) toast, roll, or pita									
Butter (not margarine)									
Margarine									

	last 4 weeks		each week			each day			
Number of times	0	1-3	1	2-4	5-6	1	2-3	4-5	6+
Vegetable soup									
Other soup									
Cornbread or tortilla									
Eggs									
Bacon									
Hot cereal, grits									
Cold cereal									
Donut									
Sweet roll or muffin									
Pancake, waffle, or French toast									
English muffin or bagel									
Biscuit									

1. What type of bread does your child usually eat:

☐ white bread ☐ whole wheat or dark bread
☐ about half and half ☐ DON'T EAT BREAD

2. What type of margarine does your child usually use:

☐ stick ☐ tub ☐ squeeze ☐ DON'T USE MARGARINE

Is this margarine:

☐ corn oil ☐ nonfat ☐ other

3. If your child eats cold breakfast cereal, what type:

☐ high fiber (All Bran) ☐ other (e.g., Corn Flakes)

4. Does your child take a multi-vitamin pill (Flintstones, TriViFlor):

☐ no ☐ yes

If yes, how often:

☐ every day ☐ 4-6 times a week ☐ 1-3 times a week ☐ less than one time a week

5. Does your child take a separate iron pill (not in the multi-vitamin pill above):

☐ no ☐ yes

6. Does your child take a separate vitamin A pill (not in the multi-vitamin pill above)"

☐ no ☐ yes

7. Does your child take a separate calcium pill (not in the multi-vitamin pill above):

☐ no ☐ yes

8. Does your child eat fried food at home:

☐ no ☐ yes

If yes, how often

☐ every day ☐ 4-6 times a week ☐ 1-3 times a week ☐ less than one time a week

If yes, what type of fat do you use to fry at home

☐ butter ☐ maragarine ☐ crisco ☐ corn oil ☐ canola oil
☐ olive oil ☐ other vegetable oil

9. Do you bake cookies, cake or pies at home:

☐ no ☐ yes

If yes, how often does your child eat home-baked cookies, cake, or pies:

☐ every day ☐ 4-6 times a week ☐ 1-3 times a week ☐ less than one time a week

If yes, what type of fat do you use to bake at home:

☐ butter ☐ maragarine ☐ crisco ☐ corn oil ☐ canola oil
☐ olive oil ☐ other vegetable oil

REFERENCES

1. Ogden CL, Carroll MD, Curtin LR, McDowell MA, Tabak CJ, Flegal KM. Prevalence of overweight and obesity in the United States, 1999-2004. *Jama.* Apr 5 2006;295(13):1549-1555.

2. Flegal KM, Carroll MD, Ogden CL, Curtin LR. Prevalence and trends in obesity among US adults, 1999-2008. *Jama.* Jan 20 2010;303(3):235-241.

3. Ogden CL, Carroll MD, McDowell MA, Flegal KM. Obesity among adults in the United States--no statistically significant change since 2003-2004. *NCHS Data Brief.* Nov 2007(1):1-8.

4. Sturm R. Increases in morbid obesity in the USA: 2000-2005. *Public Health.* Jul 2007;121(7):492-496.

5. *World Health Organization. Technical Report Series. Diet, nutrition and the prevention of chronic diseases.* Geneva 2003.

6. Luo J, Hu FB. Time trends of obesity in pre-school children in China from 1989 to 1997. *Int J Obes Relat Metab Disord.* Apr 2002;26(4):553-558.

7. Rennie KL, Jebb SA. Prevalence of obesity in Great Britain. *Obes Rev.* Feb 2005;6(1):11-12.

8. Flood A, Rastogi T, Wirfalt E, et al. Dietary patterns as identified by factor analysis and colorectal cancer among middle-aged Americans. *Am J Clin Nutr.* Jul 2008;88(1):176-184.

9. Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Actual causes of death in the United States, 2000. *Jama.* Mar 10 2004;291(10):1238-1245.

10. Finkelstein EA, Trogdon JG, Cohen JW, Dietz W. Annual medical spending attributable to obesity: payer-and service-specific estimates. *Health Aff (Millwood).* Sep-Oct 2009;28(5):w822-831.

11. Raebel MA, Malone DC, Conner DA, Xu S, Porter JA, Lanty FA. Health services use and health care costs of obese and nonobese individuals. *Arch Intern Med.* Oct 25 2004;164(19):2135-2140.

12. Burton WN, Conti DJ, Chen CY, Schultz AB, Edington DW. The role of health risk factors and disease on worker productivity. *J Occup Environ Med.* Oct 1999;41(10):863-877.

13. Bowman SA, Gortmaker SL, Ebbeling CB, Pereira MA, Ludwig DS. Effects of fast-food consumption on energy intake and diet quality among children in a national household survey. *Pediatrics.* Jan 2004;113(1 Pt 1):112-118.

14. Crespo CJ, Smit E, Troiano RP, Bartlett SJ, Macera CA, Andersen RE. Television watching, energy intake, and obesity in US children: results from the third National Health and Nutrition Examination Survey, 1988-1994. *Arch Pediatr Adolesc Med.* Mar 2001;155(3):360-365.

15. Troiano RP, Briefel RR, Carroll MD, Bialostosky K. Energy and fat intakes of children and adolescents in the united states: data from the national health and nutrition examination surveys. *Am J Clin Nutr.* Nov 2000;72(5 Suppl):1343S-1353S.

16. Lee RD, Nieman DC. *Nutritional Assessment*. fourth ed. New York City: McGraw Hill; 2007.

17. French SA, Shimotsu ST, Wall M, Gerlach AF. Capturing the spectrum of household food and beverage purchasing behavior: a review. *J Am Diet Assoc.* Dec 2008;108(12):2051-2058.

18. Baer HJ, Blum RE, Rockett HR, et al. Use of a food frequency questionnaire in American Indian and Caucasian pregnant women: a validation study. *BMC Public Health.* 2005;5:135.

19. Blum RE, Wei EK, Rockett HR, et al. Validation of a food frequency questionnaire in Native American and Caucasian children 1 to 5 years of age. *Matern Child Health J.* Sep 1999;3(3):167-172.

20. Diet History Questionnaire, Version 1.0. National Institutes of Health, Applied Research Program, National Cancer Institutes.

21. Barrett-Connor E. Nutrition epidemiology: how do we know what they ate? *Am J Clin Nutr.* Jul 1991;54(1 Suppl):182S-187S.

22. Bryant M, Stevens J. Measurement of food availability in the home. *Nutr Rev.* Feb 2006;64(2 Pt 1):67-76.

23. Nord M, Hopwood H. Recent advances provide improved tools for measuring children's food security. *J Nutr.* Mar 2007;137(3):533-536.

24. Carlson SJ, Andrews MS, Bickel GW. Measuring food insecurity and hunger in the United States: development of a national benchmark measure and prevalence estimates. *J Nutr.* Feb 1999;129(2S Suppl):510S-516S.

25. *Diet History Questionnaire Website*. http://riskfactor.cancer.gov/DHQ/. Accessed 1 Aug 2010.

26. International Food Information Council Foundation (IFIC Foundation). *Food Insight website*. http://www.foodinsight.org/. Accessed 08 Oct 2011.

27. Kerkenbush NL, Lasome CE. The emerging role of electronic diaries in the management of diabetes mellitus. *AACN Clin Issues.* Aug 2003;14(3):371-378.

28. Dowell SA, Welch JL. Use of electronic self-monitoring for food and fluid intake: A pilot study. *Nephrol Nurs J.* May-Jun 2006;33(3):271-277.

29. Sevick MA, Stone RA, Novak M, et al. A PDA-based dietary self-monitoring intervention to reduce sodium intake in an in-center hemodialysis patient. *Patient Prefer Adherence.* 2008;2:177-184.

30. USDA. ChooseMyPlate.gov. Website] http://www.choosemyplate.gov/. Accessed March 15, 2012.

31. Weinstein JL, Phillips V, MacLeod E, Arsenault M, Ferris AM. A universal product code scanner is a feasible method of measuring household food inventory and food use patterns in low-income families. *J Am Diet Assoc.* Mar 2006;106(3):443-445.

32. Lambert N, Plumb J, Looise B, et al. Using smart card technology to monitor the eating habits of children in a school cafeteria: 1. Developing and validating the methodology. *J Hum Nutr Diet.* Aug 2005;18(4):243-254.

33. Montgomery A. Creating Micro-Marketing Pricing Strategies Using Supermarket Scanner Data. *Marketing Science.* 1997;16(4):315-337.

34. Mladenic D EW, Ziolko S. Exploratory Analysis of Retail Sales of Billions of Items. Paper presented at: Proceedings of Interface, 2001.

35. Agrawal R, Imienlinski, T., Swami, A. Mining Association Rules Between Sets of Items in Large Datasets. Paper presented at: Proceedings of the ACM SIGMOD International Conference on the Management of Data, 1993.

36. Brin S, Motwani, R., Silverstein, C. Beyond Market Baskets: Generalizing Association Rules to Correlations. Paper presented at: Proceedings of the ACM SIGMOD International Conference on Management of Data, 1997.

37. Tsai P, Chen, C. Mining interesting association rules from customer databases and transaction databases. *Information Systems.* 2004:685-696.

38. Bell DR, Lattin, J.M. Shopping Behavior and Consumer Preference for Store Price Format; Why "Large Basket" Shoppers Prefer EDLP. *Marketing Science.* 1998:66-68.

39. Van Wave TW, Decker M. Secondary analysis of a marketing research database reveals patterns in dairy product purchases over time. *J Am Diet Assoc.* Apr 2003;103(4):445-453.

40. Cullen K, Baranowski T, Watson K, et al. Food category purchases vary by household education and race/ethnicity: results from grocery receipts. *J Am Diet Assoc.* Oct 2007;107(10):1747-1752.

41. Martin SL, Howell T, Duan Y, Walters M. The feasibility and utility of grocery receipt analyses for dietary assessment. *Nutr J.* 2006;5:10.

42. French SA, Wall M, Mitchell NR, Shimotsu ST, Welsh E. Annotated receipts capture household food purchases from a broad range of sources. *Int J Behav Nutr Phys Act.* 2009;6:37.

43. Narhinen M, Berg MA, Nissinen A, Puska P. Supermarket sales data: a tool for measuring regional differences in dietary habits. *Public Health Nutr.* Sep 1999;2(3):277-282.

44. Narhinen M, Nissinen A, Puska P. Sales data of a supermarket--a tool for monitoring nutrition interventions. *Public Health Nutr.* Jun 1998;1(2):101-107.

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

46. Hamilton S, Mhurchu CN, Priest P. Food and nutrient availability in New Zealand: an analysis of supermarket sales data. *Public Health Nutr.* Dec 2007;10(12):1448-1455.

47. Radimer KL, Harvey PW. Comparison of self-report of reduced fat and salt foods with sales and supply data. *Eur J Clin Nutr.* May 1998;52(5):380-382.

48. Tin ST, Mhurchu CN, Bullen C. Supermarket sales data: feasibility and applicability in population food and nutrition monitoring. *Nutr Rev.* Jan 2007;65(1):20-30.

49. Eyles H, Jiang Y, Ni Mhurchu C. Use of household supermarket sales data to estimate nutrient intakes: a comparison with repeat 24-hour dietary recalls. *J Am Diet Assoc.* Jan;110(1):106-110.

50. Ni Mhurchu C, Blakely T, Jiang Y, Eyles HC, Rodgers A. Effects of price discounts and tailored nutrition education on supermarket purchases: a randomized controlled trial. *Am J Clin Nutr.* Mar 2010;91(3):736-747.

51. Safeway. Safeway FoodFlex. Website] https://foodflex.safeway.com/. Accessed July 15, 2010.

52. Perloff J, Denbaly M. Data needs for consumer and retail firm studies. Paper presented at: American Agricultural Economics Association Meeting, 2007; Portland, Oregon.

53. U.S. Department of Agriculture. Agricultural Research Service. 2010. USDA National Nutrient Database for Standard Reference, Release 23. Nutrient Data Laboratory Home Page, http://www.ars.usda.gov/ba/bhnrc/ndl.

54. Fyfe CL, Stewart J, Murison SD, et al. Evaluating energy intake measurement in free-living subjects: when to record and for how long? *Public Health Nutr.* Feb;13(2):172-180.

55. Baxter J, Graves K, Mullis R, Potter J. Experiences in Using Computerized Sales Data to Evaluate a Nutrition Intervention Program. *J Nutr Educ.* 1996;28:443-445.

56. Jones E. An Analysis of Consumer Food Shopping Behavior Using Supermarket Scanner Data: Differences by Income and Location. *Amer. J. Agr. Econ.* 1997;79(5):1437-1443.

57. Den Hond EM, Lesaffre EE, Kesteloot HE. Regional differences in consumption of 103 fat products in Belgium: a supermarket-chain sales approach. *J Am Coll Nutr.* Dec 1995;14(6):621-627.

58. Mathios A. The importance of nutrition labeling and health claim regulation on product choice: an analysis of the cooking oils market. *Agricultural and Resource Economics Review.* 1998;27:159-168.

59. Mozaffarian D, Ludwig DS. Dietary guidelines in the 21st century--a time for food. *Jama.* Aug 11 2010;304(6):681-682.

60. Willett WC, Ludwig DS. The 2010 dietary guidelines--the best recipe for health? *N Engl J Med.* Oct 27 2011;365(17):1563-1565.

61. Bray GA, Nielsen SJ, Popkin BM. Consumption of high-fructose corn syrup in beverages may play a role in the epidemic of obesity. *Am J Clin Nutr.* Apr 2004;79(4):537-543.

62. Eyles H, Jiang Y, Ni Mhurchu C, . Use of household supermarket sales data to estimate nutrient intakes: a comparison with repeat 24-hour dietary recalls. *J Am Diet Assoc.* Jan 2010;110(1):106-110.

63. Ajani UA, Willett WC, Seddon JM. Reproducibility of a food frequency questionnaire for use in ocular research. Eye Disease Case-Control Study Group. *Invest Ophthalmol Vis Sci.* May 1994;35(6):2725-2733.

64. Ahn Y, Kwon E, Shim JE, et al. Validation and reproducibility of food frequency questionnaire for Korean genome epidemiologic study. *Eur J Clin Nutr.* Dec 2007;61(12):1435-1441.

65. Katsouyanni K, Rimm EB, Gnardellis C, Trichopoulos D, Polychronopoulos E, Trichopoulou A. Reproducibility and relative validity of an extensive semi-quantitative food frequency questionnaire using dietary records and biochemical markers among Greek schoolteachers. *Int J Epidemiol.* 1997;26 Suppl 1:S118-127.

66. Wakai K, Egami I, Kato K, et al. A simple food frequency questionnaire for Japanese diet--Part I. Development of the questionnaire, and reproducibility and validity for food groups. *J Epidemiol.* Aug 1999;9(4):216-226.

67. Ocke MC, Bueno-de-Mesquita HB, Goddijn HE, et al. The Dutch EPIC food frequency questionnaire. I. Description of the questionnaire, and relative validity and reproducibility for food groups. *Int J Epidemiol.* 1997;26 Suppl 1:S37-48.

68. Hjartaker A, Andersen LF, Lund E. Comparison of diet measures from a food-frequency questionnaire with measures from repeated 24-hour dietary recalls. The Norwegian Women and Cancer Study. *Public Health Nutr.* Oct 2007;10(10):1094-1103.

69. Ambrosini GL, de Klerk NH, O'Sullivan TA, Beilin LJ, Oddy WH. The reliability of a food frequency questionnaire for use among adolescents. *Eur J Clin Nutr.* Oct 2009;63(10):1251-1259.

70. Ambrosini GL, van Roosbroeck SA, Mackerras D, Fritschi L, de Klerk NH, Musk AW. The reliability of ten-year dietary recall: implications for cancer research. *J Nutr.* Aug 2003;133(8):2663-2668.

71. Caygill CP, Charlett A, Hill MJ. Fat, fish, fish oil and cancer. *Br J Cancer.* Jul 1996;74(1):159-164.

72. Bachman JL, Reedy J, Subar AF, Krebs-Smith SM. Sources of food group intakes among the US population, 2001-2002. *J Am Diet Assoc.* May 2008;108(5):804-814.

73. Block G. Foods contributing to energy intake in the US: data from NHANES III and NHANES 1999-2000. *J Food Comp Anal.* 2004;17(3-4):439-447.

74. Bray GA, Popkin BM. Dietary fat intake does affect obesity! *Am J Clin Nutr.* Dec 1998;68(6):1157-1173.

75. Capps O, Jr., Cleveland L, Park J. Dietary behaviors associated with total fat and saturated fat intake. *J Am Diet Assoc.* Apr 2002;102(4):490-502, 612.

76. Centers for Disease Control and Prevention Website. http://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html. Accessed 15 April 2011.

77. Gerrior S, Juan W, Basiotis P. An easy approach to calculating estimated energy requirements. *Prev Chronic Dis.* Oct 2006;3(4):A129.

78. Hu FB, Stampfer MJ, Manson JE, et al. Dietary fat intake and the risk of coronary heart disease in women. *N Engl J Med.* Nov 20 1997;337(21):1491-1499.

79. van Dam RM, Willett WC, Rimm EB, Stampfer MJ, Hu FB. Dietary fat and meat intake in relation to risk of type 2 diabetes in men. *Diabetes Care.* Mar 2002;25(3):417-424.

80. Dietary Reference Intakes for Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein, and Amino Acids (2002/2005).

81. Hooper L, Summerbell CD, Higgins JP, et al. Dietary fat intake and prevention of cardiovascular disease: systematic review. *Bmj.* Mar 31 2001;322(7289):757-763.

82. Mozaffarian D, Micha R, Wallace S. Effects on coronary heart disease of increasing polyunsaturated fat in place of saturated fat: a systematic review and meta-analysis of randomized controlled trials. *PLoS Med.* Mar 2010;7(3):e1000252.

83. Alavanja MC, Brown CC, Swanson C, Brownson RC. Saturated fat intake and lung cancer risk among nonsmoking women in Missouri. *J Natl Cancer Inst.* Dec 1 1993;85(23):1906-1916.

84. Huncharek M, Kupelnick B. Dietary fat intake and risk of epithelial ovarian cancer: a meta-analysis of 6,689 subjects from 8 observational studies. *Nutr Cancer.* 2001;40(2):87-91.

85. Lin OS. Acquired risk factors for colorectal cancer. *Methods Mol Biol.* 2009;472:361-372.

86. Nemes S, Jonasson JM, Genell A, Steineck G. Bias in odds ratios by logistic regression modelling and sample size. *BMC Med Res Methodol.* 2009;9:56.

87. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol.* Dec 1996;49(12):1373-1379.

88. Ford ES, Mokdad AH. Fruit and vegetable consumption and diabetes mellitus incidence among U.S. adults. *Prev Med.* Jan 2001;32(1):33-39.

89. Lahti-Koski M, Pietinen P, Heliovaara M, Vartiainen E. Associations of body mass index and obesity with physical activity, food choices, alcohol intake, and

smoking in the 1982-1997 FINRISK Studies. *Am J Clin Nutr.* May 2002;75(5):809-817.

90. Gittelsohn J, Wolever TM, Harris SB, Harris-Giraldo R, Hanley AJ, Zinman B. Specific patterns of food consumption and preparation are associated with diabetes and obesity in a Native Canadian community. *J Nutr.* Mar 1998;128(3):541-547.

91. Nicklas TA, Baranowski T, Cullen KW, Berenson G. Eating patterns, dietary quality and obesity. *J Am Coll Nutr.* Dec 2001;20(6):599-608.

92. Van Duyn MA, Pivonka E. Overview of the health benefits of fruit and vegetable consumption for the dietetics professional: selected literature. *J Am Diet Assoc.* Dec 2000;100(12):1511-1521.

93. Joshipura KJ, Hu FB, Manson JE, et al. The effect of fruit and vegetable intake on risk for coronary heart disease. *Ann Intern Med.* Jun 19 2001;134(12):1106-1114.

94. Mokdad AH, Bowman BA, Ford ES, Vinicor F, Marks JS, Koplan JP. The continuing epidemics of obesity and diabetes in the United States. *Jama.* Sep 12 2001;286(10):1195-1200.

95. Ford ES, Williamson DF, Liu S. Weight change and diabetes incidence: findings from a national cohort of US adults. *Am J Epidemiol.* Aug 1 1997;146(3):214-222.

96. Resnick HE, Valsania P, Halter JB, Lin X. Relation of weight gain and weight loss on subsequent diabetes risk in overweight adults. *J Epidemiol Community Health.* Aug 2000;54(8):596-602.

97. Daviglus ML, Stamler J, Orencia AJ, et al. Fish consumption and the 30-year risk of fatal myocardial infarction. *N Engl J Med.* Apr 10 1997;336(15):1046-1053.

98. Hu FB, Bronner L, Willett WC, et al. Fish and omega-3 fatty acid intake and risk of coronary heart disease in women. *Jama.* Apr 10 2002;287(14):1815-1821.

99. Kromhout D, Bosschieter EB, de Lezenne Coulander C. The inverse relation between fish consumption and 20-year mortality from coronary heart disease. *N Engl J Med.* May 9 1985;312(19):1205-1209.

100. Cotton PA, Subar AF, Friday JE, Cook A. Dietary sources of nutrients among US adults, 1994 to 1996. *J Am Diet Assoc.* Jun 2004;104(6):921-930.

101. Goodman MT, Hankin JH, Wilkens LR, Kolonel LN. High-fat foods and the risk of lung cancer. *Epidemiology.* Jul 1992;3(4):288-299.

102. Giovannucci E, Rimm EB, Stampfer MJ, Colditz GA, Ascherio A, Willett WC. Intake of fat, meat, and fiber in relation to risk of colon cancer in men. *Cancer Res.* May 1 1994;54(9):2390-2397.

103. Subar AF, Krebs-Smith SM, Cook A, Kahle LL. Dietary sources of nutrients among US children, 1989-1991. *Pediatrics.* Oct 1998;102(4 Pt 1):913-923.

104. Lin BH, Morrison RM. Higher fruit consumption linked with lower body mass index. *Food review.* 2003;25(3):28-32.

105. Cavadini C, Siega-Riz AM, Popkin BM. US adolescent food intake trends from 1965 to 1996. *Arch Dis Child.* Jul 2000;83(1):18-24.

106. Fung TT, Schulze M, Manson JE, Willett WC, Hu FB. Dietary patterns, meat intake, and the risk of type 2 diabetes in women. *Arch Intern Med.* Nov 8 2004;164(20):2235-2240.

107. Key TJ, Schatzkin A, Willett WC, Allen NE, Spencer EA, Travis RC. Diet, nutrition and the prevention of cancer. *Public Health Nutr.* Feb 2004;7(1A):187-200.

108. Glanz K, Yaroch AL. Strategies for increasing fruit and vegetable intake in grocery stores and communities: policy, pricing, and environmental change. *Prev Med.* Sep 2004;39 Suppl 2:S75-80.

109. Claro RM, Carmo HC, Machado FM, Monteiro CA. Income, food prices, and participation of fruit and vegetables in the diet. *Rev Saude Publica.* Aug 2007;41(4):557-564.

110. Epstein LH, Dearing KK, Paluch RA, Roemmich JN, Cho D. Price and maternal obesity influence purchasing of low- and high-energy-dense foods. *Am J Clin Nutr.* Oct 2007;86(4):914-922.

111. Claro RM, Monteiro CA. Family income, food prices, and household purchases of fruits and vegetables in Brazil. *Rev Saude Publica.* Dec 2010;44(6):1014-1020.

112. Cassady D, Jetter KM, Culp J. Is price a barrier to eating more fruits and vegetables for low-income families? *J Am Diet Assoc.* Nov 2007;107(11):1909-1915.

113. Mushi-Brunt C, Haire-Joshu D, Elliott M. Food spending behaviors and perceptions are associated with fruit and vegetable intake among parents and their preadolescent children. *J Nutr Educ Behav.* Jan-Feb 2007;39(1):26-30.

114. Eyles H, Rodgers A, Ni Mhurchu C. Use of electronic sales data to tailor nutrition education resources for an ethnically diverse population. *J Hum Nutr Diet.* Feb 2010;23(1):38-47.

115. Zimmermann LJ, Thompson JA, Persell SD. Electronic health record identification of prediabetes and an assessment of unmet counselling needs. *J Eval Clin Pract.* Jun 20 2011.

116. Yasnoff WA, O'Carroll PW, Koo D, Linkins RW, Kilbourne EM. Public health informatics: improving and transforming public health in the information age. *J Public Health Manag Pract.* Nov 2000;6(6):67-75.

117. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, [2007-2008], [http://www.cdc.gov/nchs/nhanes/nhanes_questionnaires.htm].

118. RODS Laboratory. *Real-Time Outbreak and Disease Surveillance web site*. https://www.rods.pitt.edu/site/. Accessed 26 Sept 2011. .

119. Hogan WR, Tsui FC, Ivanov O, et al. Detection of pediatric respiratory and diarrheal outbreaks from sales of over-the-counter electrolyte products. *J Am Med Inform Assoc.* Nov-Dec 2003;10(6):555-562.

120. Friede A, Blum HL, McDonald M. Public health informatics: how information-age technology can strengthen public health. *Annu Rev Public Health.* 1995;16:239-252.