United States Department of Agriculture

Economic Research Service

Technical Bulletin Number 1952

March 2019

Linking USDA Nutrition Databases to IRI Household-Based and Store-Based Scanner Data

Andrea C. Carlson, Elina Tselepidakis Page, Thea Palmer Zimmerman, Carina E. Tornow, and Sigurd Hermansen

United States Department of Agriculture

Economic Research Service
www.ers.usda.gov

Recommended citation format for this publication:

Andrea C. Carlson, Elina Tselepidakis Page, Thea Palmer Zimmerman, Carina E. Tornow, and Sigurd Hermansen. Linking USDA Nutrition Databases to IRI Household-Based and Store-Based Scanner Data, TB-1952, U.S. Department of Agriculture, Economic Research Service, March 2019.

United States Department of Agriculture

Economic Research Service

Technical Bulletin Number 1952

March 2019

Linking USDA Nutrition Databases to IRI Household-Based and Store-Based Scanner Data

Andrea C. Carlson, Elina Tselepidakis Page, Thea Palmer Zimmerman, Carina E. Tornow, and Sigurd Hermansen

Abstract

Americans spend about half of their food budgets to purchase about two-thirds of their food from stores. USDA purchases retail and household scanner data for food and economic research covering a broad range of Federal food and nutrition topics that relate to these food-at-home purchases. Although the data contain some nutrient data, they are not sufficient to measure how well Americans follow dietary advice or what may motivate them to do so. USDA compiles extensive nutrient and food group databases to support dietary intake studies including the National Health and Nutrition Examination Survey (NHANES). In this study, we use probabilistic and semantic matching techniques to merge the scanner data with the USDA nutrient and food composition databases. As an illustration, we use the new purchase-to-plate "crosswalk"—consisting of matches and conversion factors—to estimate the overall nutritional quality of Americans' food-at-home purchases. The 2015 Healthy Eating Index (HEI-2015) score for 2013 retail scanner sales is 55 out of 100, indicating that Americans need to substantially improve the healthfulness of their grocery purchases if they wish to follow the Federal Government's *Dietary Guidelines for Americans.*

Keywords: Scanner data, IRI InfoScan, IRI Consumer Network, Food and Nutrient Database for Dietary Studies (FNDDS), USDA National Nutrient Database for Standard Reference (SR), Food Patterns Equivalents Database (FPED), Food Patterns Equivalents Ingredient Database (FPID), probabilistic matching, semantic matching, healthy diets, Healthy Eating Index (HEI), Dietary Guidelines for Americans

Acknowledgments

This project represents the work of many people, especially the members of the USDA Linkages working group: Mark Denbaly (ERS); Kristin Koegel (USDA, Center for Nutrition Policy and Promotion (CNPP)); Kevin Kuczynski (CNPP); Biing-Hwan Lin (ERS); Mark Lino (CNPP); Carrie Martin (USDA, Agricultural Research Service (ARS)), Alanna Moshfegh (ARS), Abigail Okrent (ERS), TusaRebecca Pannucci (CNPP), Ilya Rahkovsky (ERS). Michelle Saksena (ERS) participated in the data review. Lisa Mancino (ERS), Jean Mayer (Tufts University), and Lisa Harnack (University of Minnesota) provided technical peer reviews. Thanks also to Maria Williams (ERS) and Korrin Kim (ERS) for editorial and design services.

About the Authors

Andrea Carlson and Elina Tselepidakis Page are research agricultural economists with USDA, Economic Research Service. Thea Palmer Zimmerman is a senior study manager with Westat, Inc.; Carina E. Tornow is a senior systems analyst with Westat, Inc.; Sigurd Hermansen is a project IT manager with Westat, Inc.

The analysis, findings, and conclusions expressed in this report should not be attributed to IRI.

Contents

Summary . . . . . **iv**

Introduction **1**

Data **3**

IRI Scanner Data 3

USDA Nutrient and Food Patterns Equivalents Databases 4

Differences in the Two Datasets Present Challenges for Creating a Crosswalk 4

Methods Used to Link IRI Data to USDA Nutrition Data **7**

Probabilistic and Semantic Matching 7

Data Setup 7

Establishing the Links 10

Error Rate Estimation 12

Integrity Checks 13

Issues Encountered 14

Establishing Conversion Factors **17**

Crosswalk Tables **21**

Coverage of the Purchase-to-Plate Crosswalk 22

Nutrient Data 25

Application: Measuring the Healthfulness of IRI InfoScan Purchases **27**

Conclusion **31**

References **33**

Appendix A: List of Manually Matched Linking Categories **36**

Appendix B: Sample Search Table **37**

Appendix C: The Tradeoff Between Errors and Inclusion **39**

Appendix D: Probability Proportional to Size Sampling and Estimated Error Rate **41**

Appendix E: List of Acronyms **42**

United States Department of Agriculture

A report summary from the Economic Research Service March 2019

Linking USDA Nutrition Databases to IRI Household-Based and Store-Based Scanner Data

Andrea C. Carlson, Elina Tselepidakis Page, Thea Palmer Zimmerman, Carina E. Tornow and Sigurd Hermansen

What Is the Issue?

Household- and store-based scanner data that ERS acquires from IRI are a significant resource for food economics research and policy evidence. The household-based scanner data include demographic and food purchasing information for over 120,000 U.S. households, and the store-based scanner data cover retail food sales for a large portion of the United States. However, while these data contain detailed information on purchases, prices, demographics, and stores, they are not sufficient for evaluating the healthfulness of American food purchases. To do this, the IRI scanner data need more detailed nutrient information, such as what is provided in several nutrition databases maintained by USDA. These databases keep track of the food components and nutrients of the foods most commonly consumed by Americans. They allow USDA and the U.S. Department of Health and Human Services (HHS) to assess the healthfulness of Americans' diets using the Healthy Eating Index (HEI), which measures how well diets align with USDA's Dietary Guidelines for Americans. Therefore, to expand the research capabilities of the IRI scanner data and to support USDA research on American food choices, ERS researchers in collaboration with USDA, Center for Nutrition Policy and Promotion (CNPP) and USDA, Agricultural Research Service (ARS) created a purchase-to-plate crosswalk between the IRI scanner data and USDA nutrition databases. The crosswalk allows USDA nutrition databases (nutrient and food group quantities) to be imported into the IRI data; purchase data to be attached to the USDA nutrition databases and compared to the recommendations in the Dietary Guidelines for Americans; and analysis to be conducted with the scanner data using nutrients beyond those provided by the Nutrition Facts Panel.

What Did the Study Find?

The purchase-to-plate crosswalk:

- Covers a high percentage of sales of both the 2013 IRI retail scanner data and the 2013 IRI household-based scanner data;
- Covers a total of 650,592 products in the IRI data matched to 4,390 USDA foods—representing 5.9 billion transactions in the retail data and 46.6 million transactions in the household data.

- Consists of matches between IRI food items and USDA food codes and conversion factors to convert the weight of the IRI item to the same form as the USDA nutrition databases; and
- Can be used both to import nutrients and food group data into the scanner data and to attach sales data to the USDA nutrition databases.

The linking rate—the percent of sales within a group of foods with a valid match—varies by section of the grocery store the reported food item originated from.

- The highest linking rates occur in the parts of the store where grocery items most closely resemble the foods that consumers report eating in dietary recall studies, including fresh, frozen, and canned fruits and vegetables; meat, poultry, and seafood; baked goods; condiments; snacks (shelf stable and frozen); frozen baked goods; coffee and tea; and carbonated beverages.
- Lower linking rates occur for items that consumers typically include as an ingredient in cooking (such as baking mixes), as well as for food and beverage groups that include a vast number of options, including many varieties of frozen and refrigerated meals, as well as many different flavors of mixed fruit juices and drinks. These products are less likely to have a code in the USDA nutrient databases. Low linking rates also occur for products with low sales because the study prioritized products with high volume sales.
- Because IRI food items are more granular than USDA food items (which represent an average over several products), researchers will need to exercise caution when the research question focuses on variations between closely related IRI products.

Using the crosswalk, ERS researchers estimated the HEI-2015 score for all sales in the 2013 IRI store-based scanner data to be 55 (out of 100 points), suggesting substantial room for improvement in the healthfulness of consumers' retail food purchases. (A maximum score of 100 indicates alignment or concordance with the *Dietary Guidelines for Americans*.)

How Was the Study Conducted?

Creating the crosswalk was complicated because the IRI scanner data and the USDA nutrient databases describe food differently. The IRI scanner data provide a very granular picture of the foods Americans purchase from stores. The reported food items are at the product barcode or Universal Product Code (UPC) level. Two packages of the same food product can have different UPCs if the two packages are of different sizes, flavors, package types, or sold by different retailers. On the other hand, the USDA nutrient databases use a single code to represent similar foods such as barbeque sauce or a cheese and bean burrito. Additionally, for many foods, the USDA nutrient databases provide the nutrients per gram of foods that are already prepared and cooked, rather than the purchased form. For example, squash is peeled and cooked with seeds removed, chicken is deboned, and eggs do not have shells. These differences require a set of conversion factors.

Researchers used a combination of semantic, probabilistic, and manual matching techniques to establish a purchase-to-plate crosswalk between the 2013 IRI scanner data and the 2011-12 USDA nutrient databases, the latest versions available at the time the project began. Semantic and probabilistic matching used the text descriptions from each database to identify the most likely match. Westat and USDA nutritionists reviewed the matches to improve the level of accuracy. If the automated semantic and probabilistic linking processes did not work, then researchers manually linked the products. The researchers drew the conversion factors from USDA databases, published food yield data from USDA, and in a few cases, from the product websites.

Linking USDA Nutrition Databases to IRI Household-Based and Store-Based Scanner Data

Introduction

Since 1980, Americans have had access to the healthy-eating resource, *Dietary Guidelines for Americans*, which is jointly reissued every 5 years by USDA and the U.S. Department of Health and Human Services (HHS). Since 1995, the Federal Government (first, USDA and then USDA and HHS jointly) has measured diet "quality," or healthfulness, of American diets, using the Healthy Eating Index (HEI) and dietary recall data from two linked surveys. The surveys are the National Health and Nutrition Examination Survey (NHANES) and its dietary component, What We Eat in America (WWEIA), which together include two 24-hour recalls, but do not contain food prices. Survey participants report the types and quantities of all the foods and beverages they ate and drank over a 24-hour period. Household and food retail store scanner data do contain food prices as well as more granular product information, collected over a longer period of time. Scanner data also allow researchers to construct food environment data. Together with NHANES/WWEIA, scanner data could be used to measure how well food purchases align with the *Dietary Guidelines for Americans*.

USDA purchases proprietary household and retail scanner data from IRI, a market research company, to conduct food economics research, particularly research that supports USDA's strategic goal of providing all Americans access to a safe, nutritious, and secure food supply. The scanner data include both a detailed account of the food purchases of a nationwide panel of households (the IRI Consumer Network) and store-level food sales data covering a large portion of the United States (IRI InfoScan). The IRI data include the health claims made on the package and the nutrient information that appears on the Nutrition Facts Panel (six nutrients and food energy) for many universal product code (UPC) items; these are specific to each item at the UPC barcode level.

However, the nutrient information in scanner data is not sufficient to evaluate how well household purchases and store sales align with the recommendations in the *Dietary Guidelines for Americans* or to conduct studies on the overall healthfulness of American food purchases.

To expand the use of IRI scanner data for food and economic research and monitoring, USDA established a crosswalk between the IRI data and USDA nutrition data.[1] Because the IRI data are created for marketing purposes whereas the USDA data are created for nutrition research, developing the crosswalk raised three major issues. First, the food items in the IRI data are more detailed than the more general foods in the USDA data, resulting in significantly more food items in the IRI data than in the USDA data—about 850,000 versus 7,600, respectively. Second, the naming and data organization conventions are different: whereas the IRI data uses whatever name the manufacturer uses, along with product information collected by IRI and the product's location in the

[1] In this report, we use "USDA nutrition databases" to refer to the Food and Nutrient Database for Dietary Studies (FNDDS), the Food Patterns Equivalents Database (FPED), the Food Patterns Equivalents Ingredient Database (FPID), and the National Nutrient Database for Standard Reference (SR).

grocery store, USDA uses general food types and aligns foods that are nominally and nutritionally similar. Finally, many foods in the USDA data are reported in the form that survey respondents eat them, while the IRI foods are in the form of which the item is sold.

USDA, Economic Research Service (ERS) led the project in partnership with USDA, Center for Nutrition Policy and Promotion (CNPP) and USDA, Agricultural Research Service (ARS). This report covers the links and conversion factors between the 2013 IRI scanner data and the 2011-12 USDA nutrient and food composition databases. The links apply both to IRI's store-based sales data, InfoScan, and to IRI's household panel data, the Consumer Network; however, the focus is on InfoScan. Westat, Inc. constructed the links and conversion factors under contract to ERS. Users should direct questions or comments to ERS.

Data

In order to use the IRI data for nutrition research, such as research using the HEI, we needed to append USDA nutrition data to the IRI data. The HEI requires both the quantities of each food group, sodium, and added sugars consumed, and the types of fatty acids. The links were established between the IRI product dictionaries (PDs) and the USDA Food and Nutrient Database for Dietary Studies (FNDDS). We used the food pattern databases, the Food Patterns Equivalents Database (FPED) and the Food Patterns Equivalents Ingredient Database (FPID), to measure the overall healthfulness of store purchases.

IRI Scanner Data

The IRI data purchased by USDA consist of both store-based scanner data (InfoScan) and household-level scanner data (Consumer Network). The statistical properties of both datasets are discussed in a series of ERS reports (Levin et al., 2018; Muth et al., 2016; Sweitzer et al., 2017). InfoScan provides weekly transaction data for retail food outlets such as grocery stores, club stores, convenience stores, and supercenters. Data include total sales and quantity sold at the item level. The Consumer Network provides demographic and purchase data for a nationwide panel of households, including the prices paid by the household. Both datasets record transactions at the product level, and products are identified by a barcode number, also known as the Universal Product Code (UPC).[2]

IRI's primary customers are food manufacturers and retail stores, which use the data for marketing purposes. The process of establishing links between the IRI data and the USDA nutrition databases is complicated because the UPCs are very specific. For example, two packages of the same food might have two different UPCs if the two packages are different sizes, have different flavors, or are packaged with different materials. Even if the two packages are exactly the same, the items may still have different UPCs if they were purchased at different stores.

Most UPCs in both InfoScan and the Consumer Network are contained in the same set of product dictionaries (see table 1). For this study, we matched the USDA nutrient databases to the point-of-sale product dictionary (PD POS) and perishables product dictionaries. IRI and ERS separate—into two distinct dictionaries with different product attributes—products that can be packaged in the store by the consumer and those that are prepackaged by the manufacturer. While IRI gathers product health claims and nutrition information from the package label, this information is not sufficiently populated in the IRI PDs to use in the matching process. We used the nutrition information in IRI PDs to evaluate the appropriateness of the matches.

[2] The barcode is represented by the symbol that store employees or consumers scan at the checkout counter.

Table 1

IRI product dictionaries used in this study

Name	Content	Product information
Product_dictionary_POS (POS)	Prepackaged goods other than produce	Detailed information about each UPC such as the: • Item description • Brand • Flavor • Form • Type • Manufacturer • Package size
Product_dictionary_perishables (perishables)	Products packaged by consumers or the store: • Fresh meat and seafood • Deli counter items • Bakery items • Produce including most prepackaged produce	Information about each UPC such as the: • Product • Variety • Package type • Package size (prepackaged produce items)

Note: POS = point of sale. UPC = Universal Product Code.
Source: Compiled by USDA, Economic Research Service.

USDA Nutrient and Food Patterns Equivalents Databases

The USDA nutrition databases were developed to allow researchers and policymakers to monitor the diet quality of Americans. Researchers measure diet quality by comparing the nutrient or food group[3] quantities to the recommended amounts. The primary USDA nutrition database for the matches is the Food and Nutrient Database for Dietary Studies (FNDDS) (Martin et al., 2014). FNDDS is the nutrient database for foods reported in USDA's What We Eat in America, the dietary intake component of the National Health and Nutrition Examination Survey (NHANES) (HHS, 2014). FNDDS, in turn, draws nutrient values from USDA's National Nutrient Database for Standard Reference (SR) (USDA, 2013).

The quantities for each food group are specified in two USDA databases: the Food Patterns Equivalents Database (FPED) and the Food Patterns Equivalents Ingredient Database (FPID) (Bowman et al., 2014). FPED and FPID convert foods to the 37 food groups and subgroups used in the *Dietary Guidelines for Americans* (DGA), thus allowing researchers to determine the extent to which dietary intake aligns with key recommendations in the DGA.

Differences in the Two Datasets Present Challenges for Creating a Crosswalk

USDA developed its data tables to monitor and study the nutrient and food composition of the American diet, while IRI compiled its data for market research. The differences in purpose led to differences in design that presented challenges in creating the links and conversion factors. FNDDS and the IRI data differ in three major ways that affected the creation and verification of links

[3] The major food groups are fruits, vegetables, grains, dairy (including soy milk), and protein foods. Most groups contain subgroups such as dark green vegetables, whole grains, and plant-based proteins.

between them: (1) the number of items, (2) the structure used by both datasets, and (3) the form of the items listed. These differences are summarized in table 2.

Table 2

Differences between the IRI scanner data and USDA nutrition databases

Difference	IRI scanner data	USDA nutrition databases
Primary use	Market research	Monitor and study the healthfulness of the American diet
Number of items	899,850*	7,618 (FNDDS) 3,101 (SR)
Database structure	Multiple variables; each column has similar information for every observation	Main text description, plus additional text descriptions added as needed
Form of item (product weight)	Purchase form (weight can include both edible and non-edible parts)	Raw or cooked (weight does not include inedible parts)

*This is the total number of UPCs in the two product dictionaries listed in Table 1. Not all of these UPCs had recorded sales in 2013. FNDSS = USDA Food and Nutrient Database for Dietary Studies. SR = National Nutrient Data for Standard Reference.
Source: Compiled by USDA, Economic Research Service.

Because the number of items is significantly higher in the IRI data (899,850 in the 2013 POS and perishables product dictionaries) than in FNDDS (10,719 in the 2011-12 version), the IRI information is more specific, rather than averaged over a group, as in FNDDS. Each USDA nutrient code represents a more general food item, such as barbeque sauce or bean and cheese burrito, than each UPC, which differs from other UPCs by brand, package size, flavor, or the chain selling the product. For example, 2,231 UPCs but only two USDA codes cover barbeque sauce products—one for low sodium BBQ sauce and one for all other sauces.

Second, the structures of the two datasets are different. Ideally, item entries in the FNDDS and IRI dictionaries would contain the same product attributes and variables expressed in a standard way, with columns aligned so that paired columns have the same domain of values. Unfortunately, this is not the case. The main matching variable is the food item's text description, and the two databases use different formats and phrasing in these descriptions. Where the IRI PDs have several columns describing each item, FNDDS has one main food description with additional descriptions added over time. Information in FNDDS additional description columns is not consistent between items (or rows), and it is also inconsistent within a single column. For example, the first additional description might list brand names for item 1 and available flavors for item 2. Likewise, a second additional description might list available flavors for item 1 and brand names for item 2 such that *both* columns are used to designate *both* categories—i.e., brand names and flavors—on a randomly alternating basis.

To see the differences in data descriptions, consider the dish broccoli with cheese sauce. The FNDDS description lists the key food first, followed by preparation or additives such as "Broccoli, steamed, with cheese sauce." On the other hand, the IRI PD UPC description usually starts with the brand name and includes the package size and number of units sold in a multi-pack such as "Yum Yummer's 3 Cheese Broccoli, eight single servings." Also, the Yum Yummer's 3 Cheese Broccoli may have several UPCs because the company sells the product in different package sizes and package types (bag, microwave-ready container, dry mix, frozen dinner) and the Yum Yummer's

Food Company may use different UPCs for different retail outlets. IRI PDs include other variables that indicate if the item is a refrigerated, deli, frozen, or shelf-stable item.

Finally, the form of the food is also listed differently in the two databases. The USDA databases report the nutrients per gram in the prepared and sometimes even the cooked form, while the scanner data report the weight of the purchased form. The item weights are not directly comparable between the two databases. For example, the USDA data report the weight of the edible portion of an apple, while the purchase weight recorded in the IRI data includes the edible portion plus the core, stem, and seeds. Although FNDDS includes uncooked codes for many products, the linking database matches a raw UPC to a cooked FNDDS or SR code if an appropriate raw code is not available. These differences in form require a set of conversion factors for some products.

Methods Used to Link IRI Data to USDA Nutrition Data

While it was possible to manually link each UPC to a USDA code, automating as many of the links as possible made the creation of the purchase-to-plate crosswalk more efficient. An automated approach reduced the variability caused by two or more Westat nutritionists coding similar UPCs differently, or even caused by the same nutritionist coding similar products differently. In addition, an automated process allows for the inclusion of additional UPCs in future updates.

We used the hierarchies for both the IRI and FNDDS databases and semantic matching (Doan et al., 2004) to create a search table that listed similar descriptors from both datasets. Using probabilistic matching (Fellegi et al., 1969), we compared multiple IRI and USDA descriptor variables and identified the links that were most likely to be correct. Because FNDDS codes represent more general foods than the UPCs in the IRI product dictionaries, the links between the two databases are a one-to-many match. In a one-to-many match, a UPC in the IRI product dictionary (PD) matches to one and only one FNDDS code, but most FNDDS codes have multiple UPCs linked to them.

Probabilistic and Semantic Matching

We used semantic matching to identify possible sub-text string matches between the IRI and FNDDS data. (Semantic matching searches full-text strings in one list for words and phrases in the other list that are either identical or mean similar things.) Both automated methods and human review developed the search table that paired IRI food description terms with USDA food description terms having the same meaning. Automated methods developed draft mapping rules, and then Westat nutritionists reviewed all rules and augmented the search table by identifying phrases in the IRI text descriptions that match to FNDDS.

In probabilistic matching, a program used the search table to compare the attributes in each UPC text description and other PD information to FNDDS text descriptors. Matches between attribute values (or synonyms) added to the total similarity score, while nonmatches subtracted from the score. The similarity of the two food descriptions across a number of different attributes determined a similarity score for each possible match. The program selected IRI-FNDDS food item pairs with the highest score. Data linkage posed a challenge because the semantic diversity had to be resolved when linking the IRI product dictionary text descriptions with the USDA food and nutrient text descriptions.

Data Setup

To use the power of the semantic and probabilistic matching, the data had to be prepared. In particular, we prioritized which UPCs and USDA food codes were included, created complete text descriptions, and divided the UPCs and USDA food codes into linking categories to streamline the matching process.

Selecting USDA food codes and IRI UPCs. We limited USDA food codes to those included in FPED/FPID because importing the food group quantities into the IRI data was a stated goal of the project. These data will assist researchers in estimating the Healthy Eating Index (HEI) as well as individual food group amounts.

There are nearly 1 million UPCs in the IRI data, and it is not practical to establish a link for all of them. As a starting point, we selected UPCs with reported InfoScan sales in 2013. We chose InfoScan sales over the Consumer Network purchases because InfoScan offers a more diverse set of UPCs. Future updates will also include a more complete set of links to the Consumer Network.

Users need to be aware that there are no links or conversion factors between products in the private-label product dictionary or products in the IRI Consumer Network that do not have sales in the 2013 InfoScan (table 3). The private-label product dictionary contains product information for private-label items from a set of retailers that do not disclose detailed product information at the UPC level for their private-label products.[4] This report also does not include the UPCs listed in the Consumer Network random-weight product dictionary, which covers random-weight items that are reported by a subset of the households in the panel but are not included in InfoScan.

Table 3
IRI product dictionaries

Product dictionary	Included in crosswalk
Product_dictionary_pos_2013	UPCs with sales in pos_store_2013 or pos_rma_2013
Product_dictionary_perishables_2013	UPCs with sales in RW_store_2013 or RW_rma_2013
product_dictionary_private_label	Not included
product_dictionary_RWpanel	Not included

Note: UPC = universal product code.
Source: Compiled by USDA, Economic Research Service.

To make the matching process more efficient, we reduced the number of IRI food items by combining all UPCs with the same UPC description but different package types, sizes, or retail outlets. All other IRI variables—such as category, product, brand, and other UPC descriptors (type, flavor-scent, color, style, etc.)—were the same for UPCs combined in a single food item. For example, the UPCs for frozen "Yum Yummer's Low Fat 3 Cheese Broccoli" were not included in the same food item as frozen "Yum Yummer's 3 Cheese Broccoli," and frozen "Super Delicious Food's Low Fat 3 Cheese Broccoli" would also have a different food item because the brand is different.[5] However, all "Yum Yummer's Low Fat 3 Cheese Broccoli" UPCs were included in one item regardless of package size, package type, or retail outlet. This combination allowed us to include additional UPCs that did not have reported sales.

Complete text description. Probabilistic matching algorithms search text strings for phrases that make a match. Unfortunately, the various text string columns of the IRI and FNDDS do not contain the same information. For example, the IRI data has a column for flavor, but in FNDDS, flavor could be included in the main text description or the additional description column. The IRI data contains a column for brand, but if brand is present in FNDDS, some codes have it in the main food description, while others have brand names in one of the "additional food description" observations. To fix this problem, we created complete text descriptions by combining all available text information

[4] After completion of this study, a retailer released its private-label products at the UPC level. Although these UPCs are included in the 2013 PD POS, they are not included in the crosswalk.

[5] Although we use the brand name to combine multiple UPCs, we ignored the manufacturer name. When a brand is sold by one manufacturer to another, the name of the manufacturer changes in IRI PDs. Because UPCs can remain active for multiple years, the same brand can have different manufacturers in IRI PDs.

associated with each code. For the FNDDS text description, we included the "main food description" and all "additional food description" values for each food code. To create the complete food description for the IRI PD items, we combined the description and other information in IRI PDs. The complete description included the UPC description, aisle, category, product, type, flavor, style, and brand or private label (store brand) for some product categories. Although the health claims and other nutrition information in the IRI data cover a large percentage of the volume sales, the nutrition and health claim variables are not sufficiently populated over all UPCs to use in the matching process.

Linking categories. In a standard probabilistic match, we would link all records in the IRI and FNDDS datasets at one time (Winkler, 1993). Unfortunately, a single, all-inclusive linking program was not possible because terms used in IRI PDs can mean different things depending on where the food is in the store. For example, "flat" means a flat anchovy fillet in the canned fish section, but a flatfish, such as a flounder or a halibut, in the seafood section.

In order to increase linking efficiencies, we created linking categories to limit the number of combinations generated as possible matches. Linking categories rely on similarities in the attributes of the IRI PD and FNDDS food descriptions. The resulting categories were broad enough to allow the linking categories to include products likely to match the same FNDDS codes. For example, we combined the IRI categories for fresh, frozen, and shelf-stable vegetables because FNDDS does not always distinguish the purchase form for vegetables. Examples of linking categories include mixed dishes, pudding, Mexican, eggs, fruits, vegetables, and pasta.

For most linking categories, we used the complete text descriptions from both datasets to create the matches. For a few categories, we created new columns in FNDDS to align with the information in the IRI data. For example, in the IRI data, energy drinks are described by the caffeine content, the artificial sweeteners used, the level of carbohydrates, and the brand. To complement these columns, we parsed diet, brand, and carbs from the FNDDS food description. Table 4 shows the new columns we created for the FNDDS energy drink codes. Note that not all information is included for every code. Although this variable creation had to be done separately for each linking category, this process ultimately led to better links when a limited number of attributes determined the match.

After we developed a single-text description for the IRI data (or parsed the FNDDS text description into multiple columns) and created the linking categories, the data were ready for the linking step.

Table 4
Example of restructured data from the Food and Nutrient Database for Dietary Studies (FNDDS)

Product	Diet	Brand	Carbs	Food Code
ENERGY DRINK				95311000
ENERGY DRINK	SUGAR FREE			95312600
ENERGY DRINK		FULL THROTTLE		95310200
ENERGY DRINK		MONSTER		95310400
ENERGY DRINK		MONSTER	LO CARB	95312400
ENERGY DRINK		MOUNTAIN DEW		95310500
ENERGY DRINK	SUGAR FREE	MOUNTAIN DEW		95312500
ENERGY DRINK		NO FEAR		95310550
ENERGY DRINK	SUGAR FREE	NO FEAR		95312550
ENERGY DRINK		NO FEAR MOTHERLOAD		95310555
ENERGY DRINK		NOS		95310560
ENERGY DRINK	SUGAR FREE	NOS		95312555
ENERGY DRINK		OCEAN SPRAY		95312560
ENERGY DRINK		RED BULL		95310600
ENERGY DRINK	SUGAR FREE	RED BULL		95312600
ENERGY DRINK		ROCKSTAR		95310700
ENERGY DRINK	SUGAR FREE	ROCKSTAR		95312700
ENERGY DRINK		SOBE ENERGIZE		95310750
ENERGY DRINK		VAULT		95310800
ENERGY DRINK	SUGAR FREE	VAULT ZERO		95312800
ENERGY DRINK		XS		95312900
ENERGY DRINK		XS GOLD PLUS		95312905

Source: Compiled by the authors using data from USDA, Agricultural Research Service, FNDDS 11-12.

Establishing the Links

Any matching problems require a set of match criteria to define which matches are acceptable. In this case, we had two criteria: nutrition and price. Because the links will be used to measure the overall healthfulness of food purchases, the IRI product and the FNDDS code needed to have similar amounts of each food group, energy, fatty acids, sodium, and added sugars. In addition, the matches will be used to estimate prices for FNDDS foods, so two products matched to each other needed to have a similar price. For example, because a bread mix is mostly flour, matching bread mix to flour met the nutrition criteria, but not the price criteria. This dual match criteria added to the complexity of our matching problem, and led to more unmatched UPCs than if we had simply chosen one.

With these criteria in mind, we began using semantic matching to create an initial set of sub-text string matches for each linking category. After manual review by Westat nutritionists, the list of string matches became the draft matching rules. For linking categories where we created additional variables in FNDDS, the corresponding phrases were added to the draft mapping rules. After attempting automated methods, we used a manual match for some products and a few linking categories. We documented all decisions made while establishing links in a project decision log (see

online Appendix for this report). The decision log includes decisions made during the probabilistic and semantic match creation and the manual matches. For example, we assumed that all coffee items in IRI PDs were caffeinated unless the PD included the term "decaffeinated" in either the UPC description or in one of the various attribute fields.

Automated Methods. Probabilistic matching methods and the mapping rules created draft matches for each linking category. A typical matching program creates mapping rules that match PD descriptions to very specific phrases in FNDDS. In our case, this would have created almost as many rules as there were food items in the IRI data. For example, in our program, the phrase "PEPPER very SWEET cherry RED" and "PEPPER extra SWEET rosy RED" both match to the same FNDDS code. However, a typical matching program would require two rules (one for each phrase), a third rule for "PEPPER garden fresh SWEET RED PEPPER," and so forth. Instead, we developed more general probabilistic and semantic methods of extracting phrases from the PD by rearranging a phrase or substituting a synonym. The matching program also introduced context by allowing for "regular expressions" that defined an ordered pattern of descriptors, such as "PEPPER … SWEET … RED." The program ignored PD phrases between the two keywords and matched on the pattern.

For each linking category, we used an iterative process that included reviewing IRI and USDA food descriptions, building and refining search tables, adapting and testing linking programs, selecting review samples, and reviewing those samples for accuracy. Systematic reviews of the links began with counting the number of distinct UPCs within each linking category. Wide variations in quantity sales from one linking category to the next required special attention to smaller linking categories in terms of both number of UPCs and total sales. Within each linking category, we counted the number of UPCs linked to one and only one USDA food code and labeled them. We selected the UPCs that linked to more than one food code as ambiguous links and set them aside for review.

Manual matching process. Westat nutritionists manually matched the 125,298 UPCs that were problematic for the automated linkage process to identify a match. Seventeen linking categories were manually matched, including all UPCs in the Perishables PD. (See Appendix A for a complete list of these categories.) We prioritized the manual matches by including only food items in the top 95 percent of quantity sales within each linking category. For each IRI item, one Westat nutritionist identified either an FNDDS or an SR food code or indicated that no match was available, and this determination was assigned manually. A second nutritionist reviewed all matches.

Finalizing the linking database. There are many FNDDS and SR codes that have the same description, and the linking program would have produced duplicate links if the program considered both codes at the same time. To prevent this, we used only 8-digit FNDDS codes for the automated phase, knowing that the resulting links may not match the form of the food as purchased. After completing the linking process, we reviewed the cooked FNDDS codes linked to UPC items that were uncooked and identified an SR code for the uncooked item, if available in the set of SR codes included in FPID.

To identify gaps in the linking database, we examined the linking rate for each IRI category[6] by IRI aisle and IRI department. We calculated the percentage of UPCs and 2013 sales matched for every combination of IRI category by aisle or department and flagged combinations with high sales data but no links to USDA data. We used a combination of automated and manual linking methods to add links to UPCs for these flagged items. However, there were categories, such as frozen beverages,

[6] IRI categories are not the same as the linking categories used to create the matches.

deli prepared foods, refrigerated meals and baked goods, and frozen meals where either the product dictionary did not have sufficient information, or there was not a USDA food code.

Error Rate Estimation

Linking problems represent an important tradeoff between error rates in matching and failure to match UPCs that do have a valid match.[7] For statistical purposes, we desired an error rate of less than 5 percent for each linking category, and thus, we needed to accept more omissions. A match was considered an error if (1) a better match existed, (2) a match was made but no match existed, or (3) the food item was coded as "no match exists," but a match actually existed. The true error rate, R, was defined as the total volume sold (in grams) within a linking category that was incorrectly matched divided by the total volume sold in the linking category. Specifically, R is:

$$R = \frac{\sum_{i=1}^{N} Q_i e_i}{\sum_{i=1}^{N} Q_i} \tag{1.1}$$

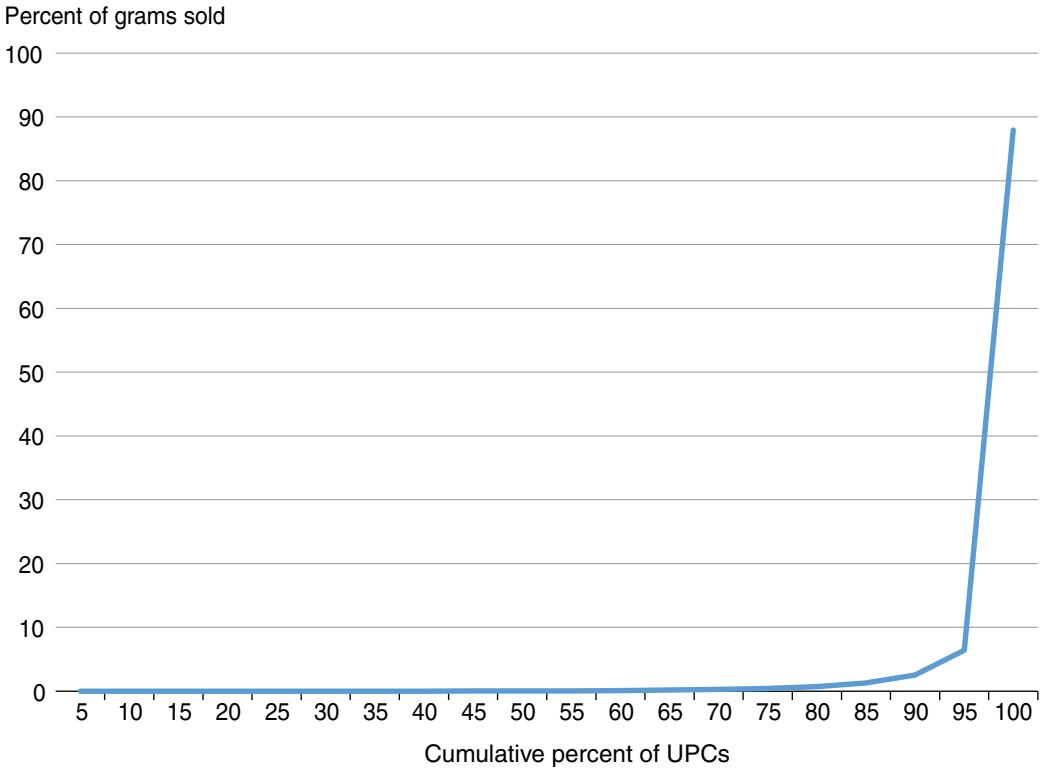
where N was the number of IRI food items in the linking category, Q was the volume sold for food item i (i = 1,...,N), $e_i = 1$ if the match for food item i was incorrect, and $e_i = 0$ if it was correctly classified. We assumed that the error rate for manually matched items and categories was very close to zero since all matches were reviewed and corrected.

It was not practical to review the matches (or nonmatches) for all 899,850 UPCs. We reduced the number by using only one UPC in a food item and applying the error designation to all UPCs in the food item. We further focused our review on UPCs with high sales volume within a linking category, since food items with the highest sales will have the greatest impact on most research outcomes. To demonstrate the magnitude of the importance of focusing on UPCs with high sales, we ranked all UPCs by the number of grams sold and divided the ranked list into 20 groups, each covering 5 percent of the UPCs. The top 5 percent of UPCs (~22,500) accounted for 88 percent of grams sold, and the top 15 percent (~67,500 UPCs) accounted for more than 95 percent of grams sold (fig. 1). Results were similar when we ranked by dollar sales instead of volume sales.

To estimate the error rate, R, for each linking category, i, in equation 1.1, we randomly selected n food items for review, focusing on the food items with the highest sales volume within the linking category. In a standard random sample, all food items would have the same probability (or chance) of being selected for the review sample. Instead, statisticians used a probability proportional to size (PPS) to create the sample (Kalton, 2014). This method still used random sampling, but the probability that a food item would be selected was based on the volume of the food item sold. Appendix D ("Probability Proportional to Size Sampling and Estimated Error Rate") shows that the error rate for the sample is a reasonable estimate for the true error rate, R.

[7] For a scientific description of the tradeoff for our problem, please see Appendix C ("The Tradeoff Between Errors and Inclusion").

Figure 1
Volume of sales were concentrated in a small share of Universal Product Codes (UPCs)



Source: Author estimates from 2013 IRI InfoScan.

Integrity Checks

Using the linking category to define sampling strata permits one to specify a sample size, n, for each linking category. To implement the sample design for each category, we ordered all food items by the volume of sales and then drew a PPS sample from the ordered list. We used SAS PROC SURVEYSELECT to draw the sample. The review sample included one UPC for each food item selected, and the "correct" or "error" designation was applied to all the UPCs associated with that food item. Westat nutritionists inspected the review sample to judge whether each link was consistent with the matching criteria. If the link was not consistent, the nutritionist indicated whether a correct match existed in the FNDDS or if the product had no acceptable match. If the estimated error rate for the category was greater than 5 percent, nutritionists revised the search table by eliminating phrases that linked on superficial text similarities and other light edits. The phrases in the search table matched either to the PD or to FNDDS. Appendix B presents an example of the search table, showing the PD phrase, the automatically matched FNDDS phrase, and the matched FNDDS phrase after manual correction.

We also compared our results to sets of matches prepared by USDA for other projects, such as the CNPP Food Prices Database (Carlson et al., 2008) for bread and cold cereal and the Cost of Fruits and Vegetables (Stewart et al., 2016) for fruits and vegetables, which we reviewed and verified prior

to use. We chose these alternative datasets because they were conducted on a smaller scale than this study and a complete review of matches was included in both projects. Fruits, vegetables, and whole grains hold particular interest for USDA researchers because Americans tend to underconsume these food groups. Because whole grain items may have lower sales than refined grain products in the same linking category, they may have been less likely to be reviewed in development. This comparison revealed that differences in matches were a result of differences in match criteria and not due to the automated match process.

To further improve the appropriateness of the links, we visually reviewed a select sample of 10,000 matches. The goal of the visual review was to identify matches that were most likely to affect research and price estimates and to identify matches that were most likely to be inconsistent with other matches. To select the links for the review sample, we created two statistics for each match: the market share and nutrient deviation. The market share was the percent of total sales of all UPCs linked to the same FNDDS code for the UPC, while the nutrient deviation score compared the nutrition facts panel data of all UPCs (when available) matched to a single FNDDS code. Because the available IRI nutrient information was more likely available for UPCs with high sales, the nutrient deviation scores were more likely available for UPCs with high sales. We estimated nutrient deviation on the following nutrients: calories, cholesterol, fiber, protein, saturated fat, sodium, total sugars, total carbohydrates, and total fat.

To ensure that the review sample was representative of all types of foods listed in the FNDDS, we selected matches in proportion to the number of items matched to FNDDS codes within each of the nine major FNDDS groups.[8] The review sample selection criteria were as follows: 40 percent of the matches had the largest market share by sales and the largest nutrient deviation; 25 percent had the largest market share, but not necessarily the largest nutrient deviation; 25 percent had the largest nutrient deviation, but not necessarily the largest market share; and the remaining 10 percent was a random sample of all other matches. The review sample market share was different from the one used to estimate error rates in a few ways. First, this review sample focused on groups created from FNDDS, rather than selecting categories within the IRI data. Second, this sample focused on the variation in key nutrients linked to a particular FNDDS code, while the review sample used for the error rate did not use the nutrient information.

USDA economists and nutritionists visually compared the IRI food description with the USDA food description to determine if the match was the best available given the match criteria, if a better match existed, or if no match existed. The latter two were considered errors. After this review, the draft matches were revised to fix all of these errors at the food item level. Due to the automated process used to establish most matches, correcting the identified matches also corrected unidentified erroneous matches of similar products.

Issues Encountered

The iterative procedures eventually produced an error rate of less than 5 percent when applied to some categories, while others required additional work to reach the desired 5-percent error rate. In some cases, these problem categories were fixed by re-categorizing food items or manually

[8] In this case, we used the first digit of FNDDS to define the groups: dairy; meat, poultry and seafood; eggs; legumes, nuts, and seeds; grains; fruit; vegetables; fats and oils; and sweets and beverages. See Martin et al. (2014) for more information on the FNDDS hierarchy.

matching poor automated matches. However, three issues remained throughout most of the study: (1) inadequate information on the UPCs in the IRI PDs, particularly the Perishables PD; (2) no matching FNDDS code for an item in the IRI PDs; and (3) implicit alternatives and missing data.

IRI collected detailed information for most packaged items presented in the data. For these items, the UPC was universal—any time an item with a given UPC was scanned, it always represented the same item.[9] However, when consumers bagged items in the produce section or the store prepared the food at the in-store deli, bakery, or meat and seafood counters, the barcode was unique to the store or retailer.

Although IRI collected data on standard produce items that were packaged by the consumer, and standard cuts of meat, poultry, and fish packaged by meat and seafood counters, some descriptions provided by the retailer were too general for the matching process—particularly in the deli and bakery sections. For example, the retailer may have used a generic description "deli platter." The deli platter sold by retailer X contained crackers, cheese, meat, and grapes, while the one sold by retailer Y contained bread, vegetables, hummus, and tabbouleh. Thus, "deli platter" and similar descriptions were not sufficient to match to a single code in FNDDS.

The second issue, no appropriate food code in FNDDS, occurred when the IRI item was a combination of different codes in FNDDS or was used as an ingredient in food preparation. Examples of no appropriate food code were baking mixes, meal kits (e.g., taco dinner kits), frozen meals, and prepared lunches. Although there were some baking mixes in FNDDS, they did not represent the wide variety of mixes available on grocery shelves. Similarly, meal kits contained several shelf-stable ingredients needed to make an entrée or side dish but lacked the non-shelf-stable items such as produce, milk, or meat, poultry, or fish.

For example, the code for tacos in FNDDS contained meat, cheese, and lettuce in addition to the ingredients included in a taco meal kit. Thus, the code for tacos was not appropriate for a taco dinner kit because these extra ingredients added nutrients and changed the food composition compared to the taco dinner kit. Given the constant changes implemented by manufacturers to satisfy consumers' demand for variety, developing individual codes for frozen meals was not practical.[10] The same was also true for meal kits, prepared deli items, and Lunchables®. As a result, there were a number of food items that did not have an FNDDS food code.

The third issue was implicit alternatives and missing defaults. An implicit alternative occurred when the PD or the FNDDS descriptions included information about one attribute, such as low-sodium, but products that contained normal amounts of sodium did not explicitly state that these products had normal amounts of sodium. While it was safe to assume the absence of the phrase "low-sodium" in a description meant a normal amount of sodium, it was difficult to put this assumption into the search table. The table contains pairs of phrases that meant the same thing, not the absence of a phrase matched to the absence of a phrase. Similarly, missing defaults occurred when one or more

[9] While manufacturers do reuse the UPC numbers assigned to them, IRI keeps track of different versions of the UPC, so that each UPC in the IRI data represents a unique product across all years of IRI data.

[10] Although dietary recall studies such as NHANES use a flag to identify individual components of a frozen meal, this technique is not practical in our case. In dietary recall data, the respondent or the coder estimates the quantities of each component for the limited number of frozen meals reported. In our case, there were too many items to individually code. For more information on combination codes, please see U.S. Department of Health and Human Services (HHS), Centers for Disease Control and Prevention (CDC) (2014) in "References."

attributes in both databases stated inclusions explicitly, such as “with pecans” or “with peanuts,” but only one database included the phrase “with walnuts” or “with cashews.”

This situation may have allowed the items with walnuts or cashews to match to an item in the other dataset that did not mention nuts. For most linking categories, pattern matching and using regular expressions were satisfactory solutions to the implicit alternatives and missing defaults that left gaps in the lists of food descriptors. However, in the future, we recommend solving the problem by expanding the number of linking categories that parse some common information from long text strings to create columns with similar information in both datasets. Although this required more time to set up each linking category than would have been required if we had started with an automated semantic linkage table, we arrived at satisfactory matches in less time for the few linking categories where we used this technique.

To our knowledge, creating a set of matches between the IRI PD and the FNDDS is one of the more complicated linking problems using food data attempted with probabilistic and semantic matching. Most applications of semantic and probabilistic matching are between two datasets where the observations represent similar information but present it in slightly different ways. In our case, the foods in the two datasets were not the same form, and the data structures were different. FNDDS represents foods in the prepared form, with some raw ingredients, while the IRI PD contains lists of foods that are purchased in grocery stores. Because the FNDDS is designed for nutrition research and the IRI PD is designed for marketing research, the type of information available in the two databases is different, even if the item represented is the same. The linking process used a combination of automated and manual matches, with intermediate review by nutritionists. The final result was 650,592 UPCs matched to 4,390 FNDSS and SR codes with a 5-percent error rate for each linking category.

Establishing Conversion Factors

USDA and IRI collect and maintain their respective databases for very different purposes. USDA uses the FNDDS, SR, FPED, and FPID databases to estimate the nutrient and food group content of the foods dietary recall study participants report *eating* during the study period. On the other hand, scanner data include foods that consumers *purchase*.[11] Food companies and retail chains use scanner data to examine market trends and consumer behavior. In order to use the linking database to assess the healthfulness of purchases, we needed to apply factors to convert the weight of a product as purchased to the weight of that product as consumed (i.e., the edible portion).

For example, a Gala apple in the IRI PD was linked to a raw apple with skin in the FNDDS data. The weight in the FNDDS data included only the edible portions of the apple, while the weight in the IRI data included the seeds, stem, and core—the purchased weight. In addition, some uncooked products in the IRI PDs linked to cooked foods in the FNDDS data because a raw or uncooked version was unavailable in the USDA nutrient databases. To the extent possible, the cooked codes linked to raw products did not contain added ingredients such as fat or salt. However, fat and moisture changes from the raw product during cooking were accounted for by the conversion factor. When a raw product linked to a cooked product, the estimated nutrient content and food group data were not the purchased nutrient content and food group data, but were as close as possible given the available data. The conversion factors converted the weight in the IRI PD to the weight in FNDDS.

Although conversion or yield factors did exist in the USDA data, they were not available at the UPC level. Different purchase forms (fresh, fresh-cut, frozen, or shelf stable) were available for the same USDA food code, and each had a different conversion factor. For example, raw broccoli in the FNDDS linked to three different IRI forms: unpackaged fresh with the tough skin and stems attached, packaged pre-trimmed and chopped, and packaged just crowns. Each form had a different amount of refuse. Other foods had different levels of preparation. For example, biscuits were sold fully cooked; refrigerated, uncooked; frozen, uncooked; or frozen, fully cooked.

In this part of the study, we developed a single multiplier to convert the weight of the UPC item as purchased to an equivalent weight of the linked FNDDS food after accounting for weight changes due to cooking and/or removal of refuse. The effort was limited to only products that had InfoScan sales data reported in 2013 and an available link. The total number of UPCs with sales and valid links was 359,572.

The first step in creating the conversion factors was to establish the purchase weight in grams of the IRI product. The IRI data reported quantities in either ounces (85 percent of records), liters (14 percent), pounds (less than 1 percent), and counts (less than 1 percent). Unfortunately, IRI did not distinguish between weight ounces and fluid ounces. For most IRI categories, we were able to assign an entire category as either fluid ounce or weight ounce. For categories that contained both fluid ounces and weight ounces, we searched the text descriptions for relevant key words. After identifying the quantities that were fluid ounces, we used the FNDDS portion codes to convert from fluid ounces to grams.

[11] In contrast to most economic studies where consumption is synonymous with purchase, researchers working with food and dietary intake refer to the foods people eat as the foods they consume, and foods people buy as foods they purchase.

Many of the products sold were already in the as-consumed form, and the conversion factor for these products was 100. We used the structure of IRI PDs to make categorical assumptions on whether the items had a conversion factor equal to 100. For example, all items in the bread, milk, crackers, and ready-to-eat breakfast cereals categories had a conversion factor of 100. Conversion factors other than 100 were required for 82 IRI categories and 448 IRI category/product groups in both the POS and perishables PDs—64,463 UPCs (18 percent of the UPCs with sales). To establish the conversion factors, we identified the forms for each food using the PD. For example, the PD identified uncooked rice, rice that cooked in 3 minutes or less, and heat-and-serve rice. Each of these types of rice had a different conversion factor, but we assumed all UPCs with the same form had the same conversion factor. Since different variables in the PD identified the form, Westat nutritionists reviewed food item descriptions to ensure food items sorted into the correct group. This review identified 7,492 additional UPCs where the conversion factor was 100. We encourage researchers to review the spreadsheets provided with the data to ensure that assumptions are appropriate for their research project.[12]

Published sources were the main sources for the conversion factors. In 1975, USDA published conversion factors for every stage of preparation (Matthews et al., 1975), which is still used by large commercial kitchens and dietitians to determine purchase quantities for a particular recipe or edible quantities of purchased products. USDA updated some of these factors to calculate the nutrient content of foods in FNDDS and SR. However, SR and FNDDS may not include the conversion factors for each stage of preparation. Although our preferred sources were FNDDS and SR, we drew from other sources when necessary. (See box "Sources of Conversion Factor Data.") Table 5 lists the number of UPCs obtained from each source. We used the IRI product dictionary information to automate the first assignment of conversion factors from these sources, followed by a Westat nutritionist's review. During this review process, nutritionists identified additional UPCs that did not require a conversion factor. The conversion factor table contains both the IRI weight in grams and the conversion factor.

Review of conversion factors. In addition to an internal review of the conversion factors by USDA nutritionists, three members of the nutrition research community who are knowledgeable about nutrient databases, the food marketplace, food yields, and nutrition policy research formed the Expert Review Panel[13] for the conversion factors. Reviewers noted that although the USDA Agriculture Handbook 102 is an older reference, a number of nutrient database developers still rely on it as a source of yield data, and thus, it was acceptable for use in this project. Reviewers agreed that the general approach for the task was a reasonable and logical method, and the assumptions about IRI forms and refuse were acceptable.

[12] The data contained in the spreadsheet include proprietary data on the product dictionary structure. These cannot be made public.

[13] Members of the Expert Review Panel were Diane Mitchell (University of Pennsylvania), Rachel Fisher (National Institutes of Health), and Sharon Kirkpatrick (University of Waterloo).

Table 5

Sources of conversion factors

Source	Count of UPCs
Conversion factor = 100	295,109
Primary sources	
FNDDS	15,797
SR	20,263
Secondary sources	
FICRCD	2,508
AH102	15,139
AH102+FICRCD	1
CNAM	206
Limited-use source	
Market check	10,549
Total	359,572

Note: FNDDS = Food and Nutrient Database for Dietary Studies. SR = National Nutrient Data for Standard Reference. FICRCD = Food Intakes Converted to Retail Commodities Database. AH102 = Agriculture Handbook No. 102. CNAM = Child Nutrition Analysis and Modeling.

Source: Author estimates using IRI PDs and data sources listed in table.

Sources of Conversion Factor Data

Primary Sources

Food and Nutrient Database for Dietary Studies (FNDDS): The "Portion Weight" table in FNDDS provided a number of portion codes, with descriptions, and an associated gram weight. For example, a portion code "1 oz. raw (yield after cooking)" indicated a gram (g) weight of 19. We used a simple calculation (19/28.35) to determine that the yield from 1 ounce (28.35 g) of the raw food was 67 percent. For some items reconstituted with water, the "recipe" for the food item in the FNDDSSRLinks[14] table provided the yield: if 0.75 ounces of dry mix and water had a final weight of 355 g, the yield for the original dry mix equaled (355/(.75*28.35) 1,669 percent.

National Nutrient Database for Standard Reference (SR): SR provided the amount of refuse as well as a description of the refuse for a number of SR food codes. For example, hard-boiled eggs had a refuse factor equal to 12 g out of 100 g, accounting for the shell. The yield for a hard-boiled egg in the shell was therefore (100-12)/100, or 88 percent.

Secondary Sources

Food Intakes Converted to Retail Commodities Database (FICRCD): We used this database with caution, as the retail form described in the database did not always correspond to the form in the IRI (Bowman et al., 2013).

Continued—

[14] In the FNDDS 2015-16, the name of this table changed to fnddsingred.

Sources of Conversion Factor Data—continued

For example, the FICRCD converted all meat, poultry, and fish to raw, boneless products, while the weights in the IRI data often included the bones. Also, any food item with several ingredients converted the ingredients to the retail form of each ingredient, which would not provide a yield factor for the combination food. For example, the FICRCD values for lasagna provided the weight of the retail ingredients (noodles, cheese, meat, etc.) needed to make the lasagna, which was not the same as the yield for cooking frozen lasagna. However, when appropriate, the values in the FICRCD provided the weight of the retail ingredient needed to yield 100 g of the food described in the food code.

Agriculture Handbook No. 102 (AH102), Food Yields: This handbook, published in 1975, provided yields at different stages of preparation. We combined the yield for one stage with the yield at another stage by multiplying the two together. For example, an IRI product for raw lamb sold with the bone linked to an SR code for boneless, cooked lamb. AH102 listed the yield for braising lamb as 68 percent, and the yield after removing the bone as 54 percent. The combined yield was therefore 37 percent including cooking and removing the bone.

Nutrient and Food Group Analysis of USDA Foods in Five of Its Food and Nutrition Programs (Child Nutrition Analysis and Modeling (CNAM)): Using similar methods to this study, the USDA Food and Nutrition Service (Zimmerman et al., 2016) developed yield factors for a small number of foods USDA offers and delivers through Federal nutrition assistance programs.

Limited-Use Source

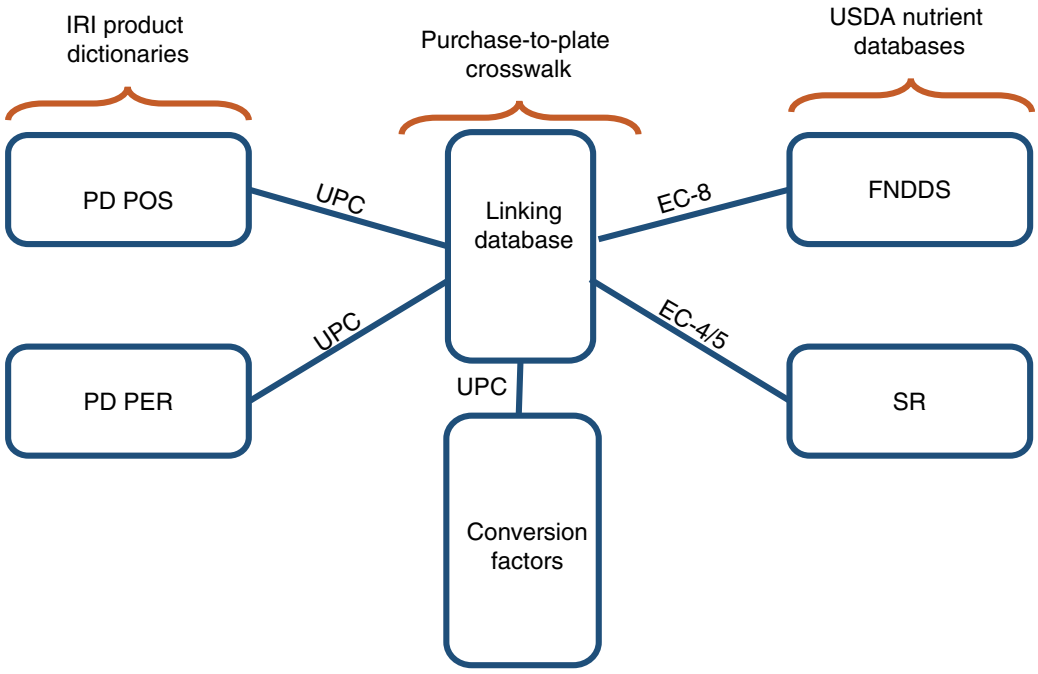
Market Check: When information was not available from any of the above sources, Westat nutritionists developed yield factors based on manufacturer information.

Crosswalk Tables

Two tables, the linking database and conversion factors table, form the purchase-to-plate crosswalk between IRI PDs and FNDDS. We used a combination of automated and manual linking methods to create the linking database. We used the structure of IRI PD to assign conversion factors from published sources to all UPCs in the linking database with 2013 InfoScan sales.

The linking database provides 899,850 matches between the 2013 IRI POS and Perishables product dictionaries. These include both the UPCs with 2013 InfoScan sales and other UPCs with the same set of attributes as those with sales. UPCs are matched to either an 8-digit FNDDS code or a 4- or 5-digit code from SR. If a match did not exist, the UPC was still included in the linking database, and a flag indicates no match is available. There are 650,592 UPCs with matches, leaving 249,258 UPCs flagged as no match available. The 359,746 UPCs included in the conversion factor table have InfoScan 2013 sales, a valid match, and sufficient information to determine the purchased form. Because the purchase-to-plate crosswalk contains proprietary information, it is available only to researchers with access to the ERS IRI data.

Figure 2
Diagram of the files



Notes: EC-8 = 8-digit FNDDS code; EC 4/5 = 4- or 5-digit NDB/SR code. PD POS = point-of-sale product dictionary. PD PER = perishables product dictionary. FNDDS = Food and Nutrient Database for Dietary Studies. UPC = Universal Product Code. SR = National Nutrient Database for Standard Reference.
Source: Constructed by USDA, Economic Research Service.

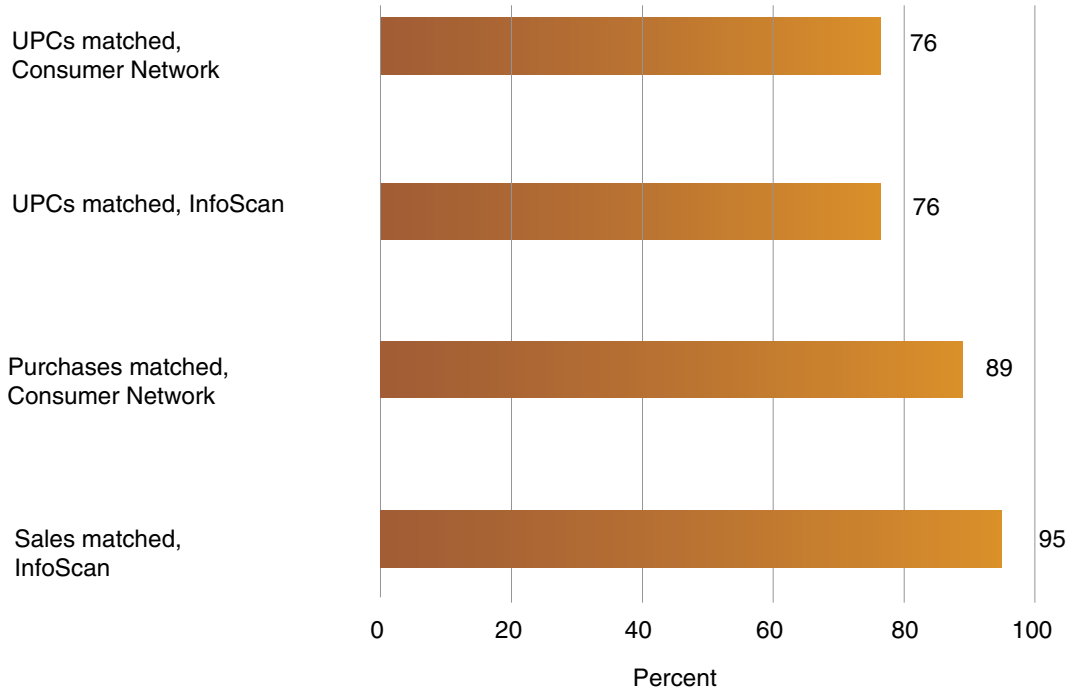
Although sample SAS code is available to users for specific applications, the tables are designed to be flexible so users can adapt them as needed. The linking database and conversion factors table join with IRI PDs and USDA nutrient databases (fig. 2). Researchers may begin by selecting UPCs of interest from the product dictionaries and then merge the UPCs with the linking database to determine the USDA food code, multiply the IRI weight purchased by the conversion factor, and finally

multiply the amount purchased by the amount of the nutrient or food component in the USDA nutrient data. Alternatively, researchers may select USDA food codes of interest, use the linking database to identify the UPCs associated with the USDA food codes, and divide the USDA weight by the conversion factor to determine expenditures and quantities of each UPC sold in retail outlets.

Coverage of the Purchase-to-Plate Crosswalk

Not all UPCs in the IRI data were included in the crosswalk. Because our goal was to provide links with an error rate of less than 5 percent for each linking category, we had to focus the crosswalk development on UPCs that covered 95 percent of total 2013 InfoScan sales.[15] Even within the list of high-sale UPCs, some products did not have a reasonable match available in the USDA nutrient databases. To provide users with a better understanding of what was and what was not included, we developed three linking rates: the percent of total UPCs matched, the percent of total sales matched, and the match rate by section of the grocery store. The relative importance of each depends on the particular research question. Figure 3 shows the overall match rate by UPCs and by sales and expenditure for both InfoScan and the Consumer Network. Because the process focused on products with sales in InfoScan, it is not surprising that the match rate by sales was higher for InfoScan than for the Consumer Network (95 versus 89 percent). The rate for the number of UPCs was about the same in the two datasets (76 percent).

Figure 3
Percent of UPCs and sales matched to USDA nutrition data by InfoScan and Consumer Network



Note: UPC = Universal Product Code.
Source: USDA, Economic Research Service estimates using 2013 IRI InfoScan and the IRI Consumer Network. IRI projection weights for the consumer panel were applied.

[15] In this case, InfoScan sales include the UPC-level transaction data in both the RMA and store files. The non-UPC private-label sales are not included.

The rate was not consistent in all areas of the grocery store (table 6). For InfoScan, the matches covered over 95 percent of sales for 52 percent of the grocery aisles,[16] and 77 percent had more than 90 percent of sales covered. For the Consumer Network, 62 percent of the aisles had coverage of at least 90 percent of purchase dollars. Note that the top five aisles with the highest InfoScan sales (dairy, liquor, produce, snacks, and fresh meat) all had match rates above 96 percent, while the four lowest sales aisles (other frozen, other refrigerated, frozen beverages, and refrigerated baked goods) had match rates below 90 percent.

Table 6

Percent of total sales matched to USDA nutrition data by aisle

Department	Aisle	Infoscan percent of sales	IRI Consumer Network percent of purchases
Perishables product dictionary			
MEAT		98.94	*
PRODUCE		96.56	54.52
SEAFOOD		98.90	*
BAKERY		93.46	*
DELI CHEESE		94.60	*
DELI MEAT		99.19	*
DELI PREPARED		84.40	*
Point of sale product dictionary			
DEPT-BEVERAGES	AISLE-CARBONATED SOFT DRINKS	99.40	98.73
DEPT-BEVERAGES	AISLE-COFFEE & TEA	95.49	95.18
DEPT-BEVERAGES	AISLE-DRINK MIXES	82.97	91.71
DEPT-BEVERAGES	AISLE-JUICES	90.75	78.06
DEPT-BEVERAGES	AISLE-NON-FRUIT DRINKS	98.14	96.95
DEPT-BEVERAGES	AISLE-SPORTS/ENERGY DRINKS	99.90	99.26
DEPT-BEVERAGES	AISLE-WATER	100.00	100.00
DEPT-FROZEN	AISLE-FROZEN BAKED GOODS	99.14	97.55
DEPT-FROZEN	AISLE-FROZEN BEVERAGES	36.38	43.05
DEPT-FROZEN	AISLE-FROZEN DESSERTS	90.23	85.01
DEPT-FROZEN	AISLE-FROZEN FRUITS & VEGETABLES	97.29	96.57
DEPT-FROZEN	AISLE-FROZEN MEALS	84.39	81.00
DEPT-FROZEN	AISLE-FROZEN MEAT/POULTRY/SEAFOOD	98.10	96.48
DEPT-FROZEN	AISLE-FROZEN SNACKS	98.18	98.28
DEPT-FROZEN	AISLE-OTHER FROZEN	83.28	91.34

[16] Although some grocery aisles have the same name as the food groups used by nutrition researchers and educators, they are not the same. For example, the grocery aisle "Dairy" includes not only milk, yogurt, and cheese (which are consistent with food groups used by nutrition researchers and educators), but also butter, margarine, whipped toppings, and other items you find in the dairy section of a grocery store (none of which are consistent with food groups used by nutrition researchers and educators).

Table 6

Percent of total sales matched by aisle—continued

Department	Aisle	Infoscan percent of sales	IRI Consumer Network percent of purchases
DEPT-GENERAL FOOD	AISLE-BABY FOOD	92.38	89.53
DEPT-GENERAL FOOD	AISLE-BAKERY	97.42	96.02
DEPT-GENERAL FOOD	AISLE-BAKING	79.23	79.96
DEPT-GENERAL FOOD	AISLE-BREAKFAST	93.89	88.59
DEPT-GENERAL FOOD	AISLE-CANDY	91.52	83.23
DEPT-GENERAL FOOD	AISLE-CONDIMENTS & SAUCES	90.26	90.75
DEPT-GENERAL FOOD	AISLE-COOKIES & CRACKERS	91.52	84.73
DEPT-GENERAL FOOD	AISLE-ETHNIC	95.14	92.99
DEPT-GENERAL FOOD	AISLE-MEALS	92.95	90.89
DEPT-GENERAL FOOD	AISLE-SNACKS	98.23	96.11
DEPT-GENERAL FOOD	AISLE-SS FRUIT	97.37	94.81
DEPT-GENERAL FOOD	AISLE-SS VEGETABLES	99.49	99.15
DEPT-LIQUOR	AISLE-LIQUOR	99.23	99.12
DEPT-REFRIGERATED	AISLE-DAIRY	99.03	99.23
DEPT-REFRIGERATED	AISLE-OTHER REFRIGERATED	77.15	69.42
DEPT-REFRIGERATED	AISLE-PRODUCE	99.55	96.98
DEPT-REFRIGERATED	AISLE-REFRIGERATED BAKED GOODS	89.54	84.09
DEPT-REFRIGERATED	AISLE-REFRIGERATED BEVERAGES	92.53	84.36
DEPT-REFRIGERATED	AISLE-REFRIGERATED CONDIMENTS	97.23	97.74
DEPT-REFRIGERATED	AISLE-REFRIGERATED DESSERTS	96.72	91.76
DEPT-REFRIGERATED	AISLE-REFRIGERATED DOUGH	77.40	77.45
DEPT-REFRIGERATED	AISLE-REFRIGERATED MEALS	59.52	60.79
DEPT-REFRIGERATED	AISLE-REFRIGERATED MEATS	95.48	95.00

*Not included in this study. Note: SS = shelf stable.
Source: USDA, Economic Research Service estimates using 2013 IRI InfoScan and the IRI Consumer Network data. IRI projection weights were used for the Consumer Network estimates.

There are three main reasons why items do not have a link: (1) the product dictionary has insufficient information; (2) a similar product does not exist in the USDA nutrient databases; or (3) the product had low or no sales in 2013. Some products in the refrigerated meals, other refrigerated meals, and deli prepared aisles have general descriptions such as “meal,” “salad,” or “sandwich,” and these descriptions are not detailed enough to match to the USDA nutrient databases. Because of the large variety of frozen meals and frozen beverages available in the market, FNDDS does not maintain codes for many frozen meals. Where possible, the frozen meals were matched to the prepared food in FNDDS, but many did not have a prepared food code available. Because the USDA databases were created to measure the nutrient content of foods eaten, FNDDS does not have codes for most baking mixes for cakes, muffins, cookies, bread, and other baked goods. Unless the mix calls only for water, we could not match to the prepared product because the prepared form has ingredients such as eggs and oil that are not included in the mix. If we had matched to the prepared item,

we might have double counted the nutrients and food components if consumers purchased the eggs and oil separately.

Nutrient Data

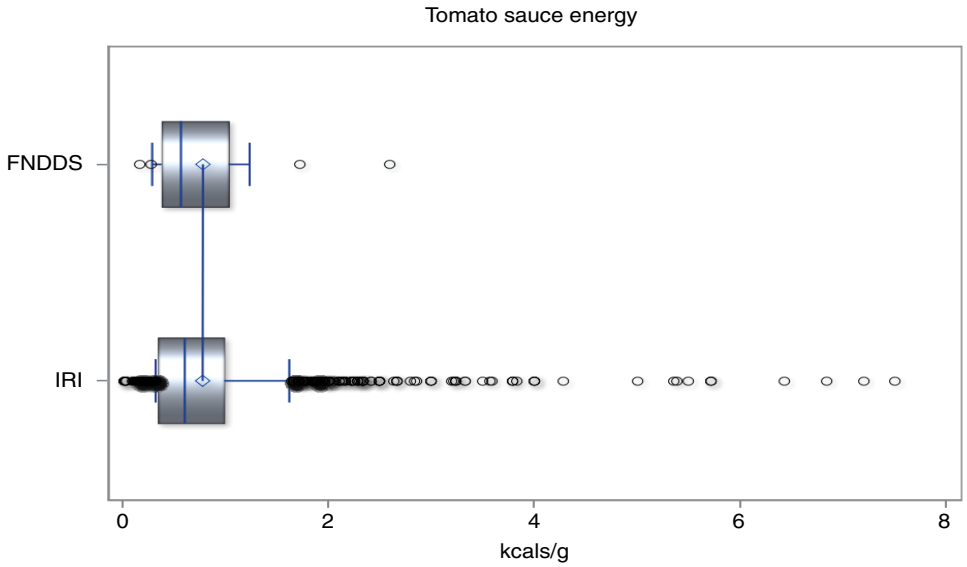
Researchers can use the linking database and conversion factors tables to determine the composition and nutrients of food sales and purchases recorded in the IRI scanner data. FNDDS contains complete nutrient information for 65 nutrients, but this information is based on averages of many products. On the other hand, the IRI data contain limited nutrient data, but the data are specific to the UPC. Although our matching criteria prioritized food groups and key nutrients such as unsaturated and saturated fats, energy, and sodium, researchers will see variation between the nutrients imported from the USDA data and the more specific IRI nutrient data.

For example, there was one FNDDS code for barbeque sauce (74406010) and 2,231 UPCs in the linking database. Of these, only 829 have nutrition data, representing 97 percent of the grams of barbeque sauce reported sold in 2013 InfoScan data. FNDDS reports that there were 1.72 kilocalories/gram (Kcals/g),[17] while the mean energy for all barbeque sauces (weighted by purchased weight) in the IRI data was 1.49 Kcals/g. However, the range was 0 to 7.5 Kcals/g. Taking this one step further, we can look at all tomato-based sauces, including barbeque sauce. In FNDDS, all tomato sauce codes began with 744. There were 21 FNDDS tomato sauce codes with reported consumption in the 2011-12 NHANES. These matched to 9,406 UPCs, but only about 4,083 have information on food energy, representing 97 percent of the grams of tomato-based sauces reported sold in the 2013 InfoScan. The box and whisker plot (fig. 4) compares the distribution of food energy (weighted by purchase or reported intake weights) in the two databases. The vertical line within each bar represents the median amount for the dataset, while the diamond represents the mean. The means of the FNDDS and IRI data are connected by a vertical line. The ends of each bar represent the 25th and 75th percentiles of energy quantities, the horizontal blue lines (the whiskers) show the range from the 5th to the 95th percentiles, and the dots represent observations outside of the 5th and 95th percentiles. Note that the two means of food energy are very close, but the IRI products have a wider range of Kcal/g than those in the FNDDS. Other nutrients included in both datasets have similar patterns.

Researchers need to be aware of these differences when choosing whether to use the more complete but general nutrients from those in the FNDDS or the UPC-specific data for their specific research question. Some research questions may require using both types of nutrition data and additional comparisons of IRI nutrient data and USDA nutrient estimations.

[17] The nutrition facts label and nonacademic discussions of food energy use the term “calorie” to refer to the kilocalorie, which equals 1,000 calories. In this report, we use the scientific kilocalorie (Kcal) to report quantities of food energy.

Figure 4
Distribution of food energy (kcals/g) of tomato sauce in the FNDDS and IRI data



Note: Distribution includes all FNDDS codes beginning with 744, and UPCs with energy data in the IRI data matched to these FNDDS codes. Mean kcal/g (diamond) is weighted by consumption (FNDDS) or purchase (IRI). The vertical line within each bar represents the median amount for the dataset, while the diamond represents the mean. The means of the FNDDS and IRI data are connected by a vertical line. The ends of each bar represent the 25th and 75th percentiles of energy quantities, the horizontal blue lines (the whiskers) show the range from the 5th to the 95th percentiles, and the dots represent observations outside of the 5th and 95th percentiles. kcals/g = kilocalories per gram. FNDDS = Food and Nutrient Database for Dietary Studies.
Source: USDA, Economic Research Service estimates using 2013 IRI InfoScan and the National Center for Health Statistics, Centers for Disease Control 2011-12 National Health and Nutrition Examination Survey (NHANES).

Application: Measuring the Healthfulness of IRI InfoScan Purchases

There are a wide variety of research questions that can be addressed with the crosswalk, including many that involve the Healthy Eating Index (HEI). (See box "The Healthy Eating Index.") Americans spend about half of their total food budgets on food from stores, but purchase about two-thirds of their calories from stores (Lin et al., 2012). Previous measurements of the diet quality of food-at-home relied on self-reported food intake data (Lin et al., 2012) or used the limited set of food advertised in store circulars (Jahns et al., 2016). These researchers used the Healthy Eating Index (HEI) (Schap et al., 2017) to measure diet quality. Because estimating the HEI score requires detailed food composition data, the score could not be directly estimated using scanner data before constructing the crosswalk. Previous research (Volpe et al., 2012; Volpe et al., 2013) imputed HEI scores for Nielsen Homescan participants using the dietary recall data in NHANES. We used InfoScan rather than the Consumer Network to estimate the HEI scores for store-based food purchases because previous ERS research indicates that the Consumer Network's survey participants have underreported certain purchases, including fruits and vegetables, compared to other surveys (Sweitzer et al., 2017). Although the InfoScan data available to ERS represent only half of all grocery sales (Levin et al., 2018), to our knowledge, there are no studies examining how nationally representative the sales are.

The Healthy Eating Index

We measure the healthfulness of food purchases using the Healthy Eating Index (HEI-2010 and HEI-2015). The Healthy Eating Index measures how well individual diets of study participants comply with the *Dietary Guidelines for Americans*. The score consists of 100 possible points, derived from 12 (HEI-2010) or 13 (HEI-2015) components (see box table). The scores are based on a universal set of standards for all age-gender groups, and the scoring standards are based on the amount of the component per 1,000 kilocalories. Note that for the adequacy components, a higher score indicates higher consumption, but for the moderation components, a higher score indicates lower consumption. In other words, a higher score always means closer compliance with the *Dietary Guidelines for Americans*. The scoring mechanism is designed to allow researchers to estimate scores at any level—from 1 day of an individual's diet to a year of the national food supply (Miller et al., 2015). Every 5 years, USDA and HHS update the *Dietary Guidelines for Americans*, and the HEI score is also updated.

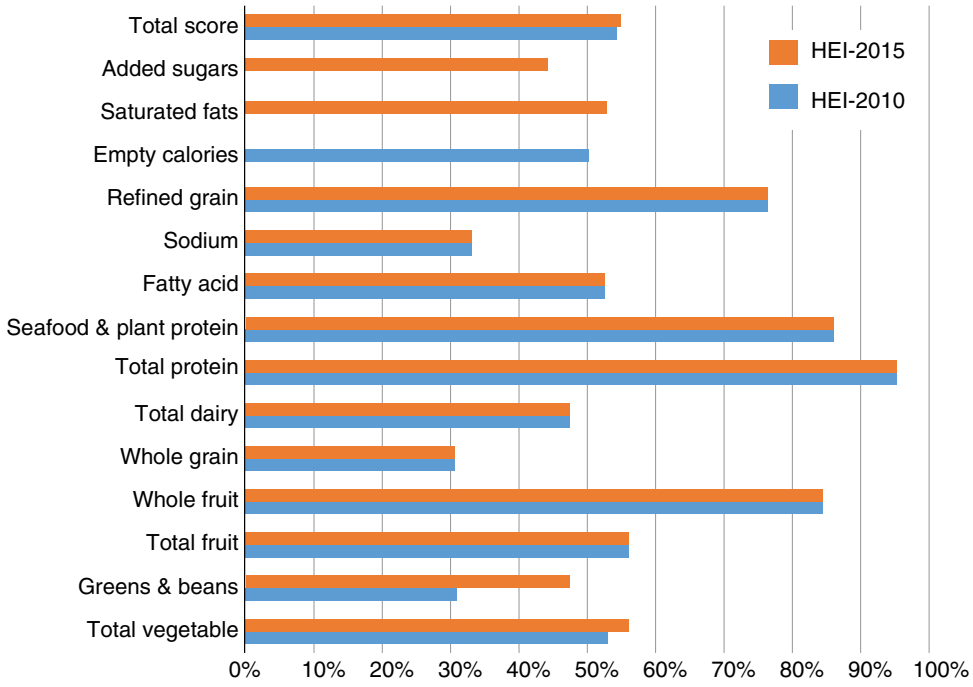
Healthy eating index (HEI) components

	Construct	Max Points	Component	Standard for Maximum Points		Standard for Minimum Score of Zero	
				2010	2015	2010	2015
Adequacy	Fruits	10	Total Fruits (5pts)	≥0.8 cup		No Fruits	
			Whole Fruits (5pts)	≥0.4 cup		No Whole Fruits	
	Vegetables	10	Total Veg. (5pts)	≥1.1 cup		No Vegetables	
			Greens & Beans (5pts)	≥0.2 cup		No Dark Green Vegetables or Legumes	
	Grains	10	Whole Grains	≥1.5 oz		No Whole Grains	
	Dairy	10	Milk/Dairy	≥1.3 cup		No Dairy	
	Protein Foods	10	Total Protein Foods (5pts)	≥2.5 oz		No Protein Foods	
			Seafood & Plant Proteins (5pts)	≥0.8 oz		No Seafood or Plant Proteins	
	Fats	10	Fatty Acid Ratio	(PUFAs + MUFAs)/SFAs ≥2.5		(PUFAs + MUFAs)/SFAs ≤1.2	
Moderation	Refined Grains	10	Refined Grains	≤1.8 oz		≥4.3 oz	
	Sodium	10	Sodium	≤1.1 gram		≥2.0 grams	
	Empty Calories	20	Empty Calories (20 pts)	≤19% of energy	---	≥ 50% of energy	---
			Added Sugars (10pts)	---	≤6.5% of energy	---	≥26% of energy
			Saturated Fats (10pts)	---	≤8% of energy	---	≥16% of energy

Notes: PUFA = polyunsaturated fatty acids. MUFA = monounsaturated fatty acids. SFA = saturated fatty acids. pts = points. veg. = vegetables. For more information on the HEI score, please see the Healthy Eating Index pages on the websites for USDA, Center for Nutrition Policy and Promotion and the National Institutes of Health.
Source: Krebs-Smith et al. (2018).

We estimated the HEI-2010 and HEI-2015 scores using the simple scoring method. That is, we added the quantity of each component (e.g., total fruit, greens, and beans) across all the food sold in InfoScan, and divided by the number of calories for all food sold in InfoScan. We did not include items sold by counts (e.g., number of bagels, muffins) because we did not have a weight for these items. IRI provides weights for produce items typically sold by count. The total (54 for HEI-2010 and 55 for HEI-2015)—as well as our estimates of both the HEI-2010 and HEI-2015 as a percent of the total possible score for each component—indicate that foods purchased in retail outlets are not aligned with key recommendations in the *Dietary Guidelines for Americans* (fig. 5). Although the total protein score exceeds 90 percent, indicating Americans are close to purchasing enough protein foods to meet recommendations, the seafood and plant protein score (4.3 out of 5 points or 86 percent of total) indicates that protein foods purchased do not include enough seafood or plant-based proteins.

Figure 5
Percent of maximum total and component Healthy Eating Index (HEI) scores, 2010 and 2015



Note: The components of added sugars and saturated fats are not in the HEI-2010 score, and the empty calories component is not in the HEI-2015 score.
Source: USDA, Economic Research Service estimates using 2013 IRI InfoScan.

Our total HEI-2010 score was comparable to estimates using other data such as food availability data (ERS Food Availability), weekly household acquisition data (FoodAPS), and dietary intake data (NHANES). The Food Availability data measured the amount of food commodities available at each stage of the food supply chain—from farm to consumers. An estimate of the HEI score for the food available to consumers covered all food available for human consumption, including food sold through retail outlets, restaurants, and other food-away-from-home sources, as well as food given away by institutions (Miller et al., 2015). Two recent ERS studies measured the HEI score of all foods acquired in a week by each of the approximately 5,000 households in the USDA, Food Acquisition and Purchase Survey (FoodAPS), including food that the household purchased (both from food-at-home retail stores and food-away-from-home retail outlets) or acquired for free. One study (Mancino et al., 2018b) compared these estimates using dietary recall data in NHANES, while the second study (Mancino et al., 2018a) broke out HEI scores by all foods acquired and foods purchased from different types of food vendors, such as large grocery stores, and food away from home. Both studies first calculated the HEI score over all foods acquired by the household (or consumed by an individual NHANES participant) and took the average of all households. Reedy and co-authors (2018) estimated population level HEI-2010 and HEI-2015 scores using NHANES. Rather than a simple mean of each individual, they estimated the HEI score by first estimating the usual intake for a group of individuals (Kirkpatrick et al., 2018).

Table 7 shows how our HEI-2010 score compared to these other estimates. Our estimate (54) was similar to the other estimates. All results indicate the need for improvement. This close result is a little surprising given that four of the five estimates included food away from home, which is

generally less healthy than the food at home represented by InfoScan (Mancino et al., 2009; Todd et al., 2010). NHANES includes foods that Americans reported eating, while the InfoScan results include both foods eaten and foods that are purchased and later thrown out. Unlike the Food Availability Data, which measure food at the commodity level, the quantities of each individual item sold, eaten, or acquired are precisely measured in InfoScan, NHANES, and FoodAPS. In addition, not all retail chains and stores choose to participate in IRI or to allow IRI to provide their data to ERS. Although more research is needed, the similarity of scores may provide evidence that the InfoScan data purchased by USDA is representative of the foods Americans purchase.

Table 7
Comparison of Healthy Eating Index (HEI)-2010 score estimates

Study	Data used	Unit	Calculation method[a]	HEI-2010
This report, TB-1952	IRI-InfoScan 2013	All retail purchases (annual)	Simple sum of all stores	54
Miller et al. (2015)	ERS Food Availability 2010	All available food (annual)	Simple sum of all available food	55
Reedy et al. (2018)	NHANES 2011-12	Population	Population ratio method	56
Mancino et al. (2018b)	NHANES 2011-12	Individual food intake (day)	Average of individuals	54
Mancino et al. (2018a)	FoodAPS 2012-13	Household acquisition (week)	Average of households	53
Mancino et al. (2018a)	FoodAPS 2012-13	Household acquisition from large grocery stores (week)	Average of households	52

[a] Information on calculation methods is available in Kirkpatrick et al. (2018).
Source: Constructed by USDA, Economic Research Service.

Conclusion

USDA purchases proprietary household and retail scanner data to conduct research on consumer behavior, food prices, available new products, and the healthfulness of consumer choices. This report contributes to the research on healthy food purchases by providing a purchase-to-plate crosswalk between the USDA nutrient and food group data and the scanner data purchased by USDA. We used a combination of semantic, probabilistic, and manual matching to establish links for approximately 95 percent of the sales recorded in the 2013 InfoScan and 89 percent of sales in the 2013 Consumer Network. We used USDA food yield data collected over several decades to convert the purchase weight to the edible weight used in the USDA nutrition databases. By using the crosswalk, we estimated an HEI-2010 score of 54 and an HEI-2015 score of 55. This estimate is close to HEI estimates obtained for household food acquisition data, dietary intake, and national food availability data and indicates that the foods purchased at the retail level are not in alignment with the *Dietary Guidelines for Americans*.

The matches and conversion factors established by methods described in this report do have a couple of limitations that researchers should consider. First, the crosswalk does not cover all products included in the scanner data purchased by USDA. The crosswalk does not cover the private-label data that are not provided to USDA at the UPC level; UPCs that are reported purchased by the Consumer Network participants but do not have recorded sales in InfoScan; and the random weight data from the Consumer Network. As noted in the first ERS statistical properties report (Muth et al., 2016), some retail chains do not release their private-label data at the UPC level. The limited information available on these products, especially package size, makes estimations of the healthfulness of purchases impossible. In 2013, the unmatched private-label products covered approximately 6.8 percent of all sales in InfoScan. Because the research focuses on UPCs with sales in InfoScan, products with sales in the Consumer Network may not be included even if a valid match exists. Of particular note are random-weight items—items that are packaged by the consumer or store and thus do not have a UPC. A subset of Consumer Network participants record expenditures, but not quantities for random-weight products. Even if the linking database included these products, researchers would not be able to include them in assessing the healthfulness of purchases. The crosswalk's second limitation is that the nutrients and food pattern equivalent data in the USDA databases are an average of many individual products. Thus, the nutrient and food composition data are not unique to each UPC.

Researchers should also consult the set of statistical properties reports released by ERS (Levin et al., 2018; Muth et al., 2016; Sweitzer et al., 2017). Of particular note for researchers wishing to estimate HEI scores for participants in the Consumer Network is the underreporting of certain categories, especially eggs, fresh vegetables, fresh fruit, fish, and seafood. Expenditures in these four categories constitute 47-50 percent of the Consumer Expenditure Survey and 45-59 percent of FoodAPS food-at-home estimates. The differences vary across demographic groups—high-income and large households tend to report a smaller share of their expenditures (Sweitzer et al., 2017). Researchers should exercise caution when examining the overall healthiness of food purchases using the Consumer Network. Similarly, although the InfoScan data released to ERS cover about half of all sales reported in the Economic Census of Food Sales at Payroll Establishments (Levin et al., 2018), the coverage varies by type of store and region of the country. Although our estimates of the Healthy Eating Index are similar to estimates using other data, researchers should exercise caution when drawing conclusions about the healthfulness of retail food purchases, since the ERS scanner data is not necessarily representative.

Despite these limitations, the purchase-to-plate crosswalk represents the most comprehensive attempt to merge retail food scanner data to the USDA nutrient and food composition databases to date. This is also a first attempt to use probabilistic and semantic methods to reduce the level of manual effort required to produce these matches. Unfortunately, the IRI data structure and text descriptions change significantly enough from year to year that we will not be able simply to run the linking programs on other years of the data to produce matches. Future updates will continue to employ a combination of automated and manual matches, with a greater focus on the data setup and using smaller linking categories. There is both a need for continued research on ways to improve the linking process and for development of more USDA food codes to cover more products that Americans purchase.

References

Bowman, S.A., C.L. Martin, J.L. Carlson, J.C. Clemens, B.H. Lin, and A.J. Moshfegh. 2013. *Food Intakes Converted to Retail Commodities Databases 2003-08: Methodology and User Guide*, U.S. Department of Agriculture, Agricultural Research Service and Economic Research Service.

Bowman, S.A., J.C. Clemens, J.E. Friday, R.C. Thoerig, and A.J. Moshfegh. 2014. *Food Patterns Equivalents Database 2011-12: Methodology and User Guide*, U.S. Department of Agriculture, Agricultural Research Service, Beltsville Human Nutrition Research Center, Food Surveys Research Group.

Carlson, A., M. Lino, W. Juan, K. Marcoe, L. Bente, H.A.B. Hiza, P.M. Guenther, and E. Leibtag. 2008. *Development of the CNPP Prices Database CNPP-22*, U.S. Department of Agriculture, Center for Nutrition Policy and Promotion.

Doan, A., N.F. Noy, and A.Y. Halevy. 2004. "Introduction to the Special Issue on Semantic Integration," SIGMOD Record 33 (4):11-13.

Fellegi, I., and A. Sunter. 1969. "A Theory for Record Linkage," J*ournal of the American Statistical Association* 64 (328):1183-1210.

Hajian-Tilaki, K. 2013. "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation," *Caspian Journal of Internal Medicine* 4 (2):627-635.

Jahns, L., A.J. Scheett, L.K. Johnson, S.M. Krebs-Smith, C.R. Payne, L.D. Whigham, B.S. Hoverson, and S. Kranz. 2016. "Diet Quality of Items Advertised in Supermarket Sales Circulars Compared to Diets of the U.S. Population, as Assessed by the Healthy Eating Index-2010," *Journal of the Academy of Nutrition and Dietetics* 116 (1):115-122 e111.

Kalton, G. 2014. "Systematic Sampling," in Wiley Statsref: Statistics Reference Online. John Wiley & Sons, Ltd.

Kirkpatrick, S.I., J. Reedy, S.M. Krebs-Smith, T.E. Pannucci, A.F. Subar, M.M. Wilson, J.L. Lerman, and J.A. Tooze. 2018. "Applications of the Healthy Eating Index for Surveillance, Epidemiology, and Intervention Research: Considerations and Caveats," *Journal of the Academy of Nutrition and Dietetics* 118 (9):1603-1621.

Krebs-Smith, S.M., T.E. Pannucci, A.F. Subar, S.I. Kirkpatrick, J.L. Lerman, J.A. Tooze, M.M. Wilson, and J. Reedy. 2018. "Update of the Healthy Eating Index: HEI-2015," *Journal of the Academy of Nutrition and Dietetics* 118 (9):1591-1602.

Levin, D., D. Noriega, C. Dicken, A. Okrent, M. Harding, and M. Lovenheim. 2018. *Examining Store Scanner Data: A Comparison of the IRI Infoscan Data With Other Data Sets, 2008-12,* TB-1949, U.S. Department of Agriculture, Economic Research Service.

Lin, B.H., and J. Guthrie. 2012. *Nutritional Quality of Food Prepared at Home and Away From Home, 1977-2008*, EIB-105, U.S. Department of Agriculture, Economic Research Service.

Mancino, L., J. Todd, and B.-H. Lin. 2009. "Separating What We Eat from Where: Measuring the Effect of Food Away From Home on Diet Quality," *Food Policy* 34 (6):557-562.

Mancino, L., J. Guthrie, M.V. Ploeg, and B.H. Lin. 2018a. *Nutritional Quality of Foods Acquired by Americans: Findings From USDA's National Household Food Acquisition and Purchase Survey*, EIB-188, U.S. Department of Agriculture, Economic Research Service.

Mancino, L., J.E. Todd, and B. Scharadin. 2018b. *USDA's National Household Food Acquisition and Purchase Survey: Methodology for Imputing Missing Quantities to Calculate Healthy Eating Index-2010 Scores and Sort Foods Into ERS Food Groups*, TB-1947, U.S. Department of Agriculture, Economic Research Service.

Martin, C.L., J.B. Montville, L.C. Steinfeldt, G. Omolewa-Tomobi, K.Y. Heendeniya, M.E. Adler, and A.J. Moshfegh. 2014. *USDA Food and Nutrient Database for Dietary Studies 2011-2012*, U.S. Department of Agriculture, Agricultural Research Service, Food Surveys Research Group.

Matthews, R.H., and Y.J. Garrison. 1975. *Food Yields Summarized by Different Stages of Preparation Handbook 102*, U.S. Department of Agriculture, Agricultural Research Service.

Miller, P.E., J. Reedy, S.I. Kirkpatrick, and S.M. Krebs-Smith. 2015. "The United States Food Supply Is Not Consistent with Dietary Guidance: Evidence From an Evaluation Using the Healthy Eating Index-2010," *Journal of the Academy of Nutrition and Dietetics* 115 (1):95-100.

Muth, M.K., M. Sweitzer, D. Brown, K. Capogrossi, S. Karns, D. Levin, A. Okrent, P. Siegel, and C. Zhen. 2016. *Understanding IRI Household-Based and Store-Based Scanner Data*, TB-1942, U.S. Department of Agriculture, Economic Research Service.

Reedy, J., J.L. Lerman, S.M. Krebs-Smith, S.I. Kirkpatrick, T.E. Pannucci, M.M. Wilson, A.F. Subar, L.L. Kahle, and J.A. Tooze. 2018. "Evaluation of the Healthy Eating Index-2015," *Journal of the Academy of Nutrition and Dietetics* 118 (9):1622-1633.

Schap, T., K. Kuczynski, and H. Hiza. 2017. "Healthy Eating Index—Beyond the Score," *Journal of the Academy of Nutrition and Dietetics* 117 (4):519-521.

Stewart, H., J. Hyman, A. Carlson, and E. Frazão. 2016. *The Cost of Satisfying Fruit and Vegetable Recommendations in the Dietary Guidelines*, EB-27, U.S. Department of Agriculture, Economic Research Service.

Sweitzer, M., D. Brown, S. Karns, M.K. Muth, P. Siegel, and C. Zhen. 2017. *Food-at-Home Expenditures: Comparing Commercial Household Scanner Data From IRI and Government Survey Data*, TB-1949, U.S. Department of Agriculture, Economic Research Service.

Todd, J.E., L. Mancino, and B.-H. Lin. 2010. *The Impact of Food Away From Home on Adult Diet Quality*, ERR-90, U.S. Department of Agriculture, Economic Research Service.

U.S. Department of Agriculture, Agricultural Research Service, Nutrient Data Laboratory. 2013. *Nutrient Database for Standard Reference, Release 26*.

U.S. Department of Health and Human Services (HHS), Centers for Disease Control and Prevention (CDC). 2014. *National Health and Nutrition Examination Survey 2011-12*.

U.S. Department of Health and Human Services (HHS), and U.S. Department of Agriculture (USDA). 2015. *2015-2020 Dietary Guidelines for Americans: 8th Edition*, U.S. Government Printing Office.

Volpe, R., and A. Okrent. 2012. *Assessing the Healthfulness of Consumers' Grocery Purchases*, EIB-102, U.S. Department of Agriculture, Economic Research Service.

Volpe, R., A. Okrent, and E. Leibtag. 2013. "The Effect of Supercenter-Format Stores on the Healthfulness of Consumers' Grocery Purchases," *American Journal of Agricultural Economics* 95 (3):568-589.

Winkler, W.E. 1993. *Improved Decision Rules in the Fellegi-Sunter Model of Record Linkages*, U.S. Department of Commerce, Census Bureau, Center for Statistical Research and Methodology.

Zimmerman, T.P., B. Sun, and S. Dixit-Joshi. 2016. *Nutrient and Food Group Analysis of USDA Foods in Five of Its Food and Nutrition Programs-2014*, U.S. Department of Agriculture, Food and Nutrition Service.

Appendix A: List of Manually Matched Linking Categories

Table A
List of manually matched linking categories

Linking category	IRI categories
ALCX	COCKTAIL MIXES
BREX	BAKING MIXES
CAKE	BAKED GOODS - RFG; BAKERY SNACKS; BAKING MIXES; CHEESECAKES - RFG; DESSERTS/TOPPINGS - FZ; PIES & CAKES
CANDY	BAKING NEEDS; BREATH FRESHENERS; CHOCOLATE CANDY; COUGH DROPS; DRY FRUIT SNACKS; GUM; MARSHMALLOWS; NON-CHOCOLATE CANDY; OTHER SNACKS
CEREAL	COLD CEREAL; OTHER BREAKFAST FOOD - SS
COFX	COFFEE, IRI product = COFFEE ADDITIVE/FLAVORING
COOKIE	BAKING MIXES; BAKING NEEDS; BREAD/DOUGH - FZ; COOKIES; DOUGH/ BISCUIT DOUGH - RFG
CRACKER	CRACKERS; MATZOH FOOD; RICE/POPCORN CAKES
DIPX	DIP/DIP MIXES - SS
GRAX	GRAVY/SAUCE MIXES
ICECREAM	ICE CREAM CONES/MIXES; ICE CREAM/SHERBET; NOVELTIES - FZ
INGR	BAKING NEEDS; OTHER CONDIMENTS - RFG; OTHER FOODS - FZ; SPICES/ SEASONINGS; VINEGAR
JUICE	ASEPTIC JUICES; BABY FORMULA/ELECTROLYTES; BOTTLED JUICES - SS; CANNED JUICES - SS; DRINK MIXES; JUICES - FZ; JUICES/DRINKS - RFG; SEAFOOD - SS
MXD	ASIAN FOOD; DINNERS - SS; DINNERS/ENTREES - FZ; ENTREES - RFG; LUNCHES - RFG
PBD	OTHER BREAKFAST FOOD - SS; WEIGHT CONTROL
RECP	DRY PACKAGED DINNER MIXES; OTHER BREAKFAST FOOD - SS
PERISHABLES	PERISHABLES PRODUCT DICTIONARY

Source: Compiled by the authors using the IRI product dictionaries.

Appendix B: Sample Search Table

Table B
Sample search table

PD phrase	Automated FNDDS match	Corrected FNDDS match
AMBROSIAL GRANOLA, GRANOLA	NFS, GRANOLA	
AMPORT FOODS, GRANOLA	NFS, GRANOLA	
BAKERY ON MAIN, GRANOLA	HOMEMADE, GRANOLA	NFS, GRANOLA
BEAR NAKED, GRANOLA	HOMEMADE, GRANOLA	NFS, GRANOLA
BREAD & CIE, GRANOLA	HOMEMADE, GRANOLA	NFS, GRANOLA
HEALTH VALLEY FIBER 7, 7 GRAIN	HEALTH VALLEY, FIBER 7 FLAKES	
HEALTH VALLEY FIBER 7, ORGANIC MULTIGRAIN	HEALTH VALLEY, FIBER 7 FLAKES	
HEALTH VALLEY OAT BRAN OS, OAT BRAN	HEALTH VALLEY, OAT BRAN FLAKES	
HEALTH VALLEY RAISIN BRAN, BRAN	HEALTH VALLEY, OAT BRAN FLAKES	NFS, RAISIN BRAN
HEALTH VALLEY, AMARANTH	HEALTH VALLEY, OAT BRAN FLAKES	AMARANTH
HEALTH VALLEY, GRANOLA	HEALTH VALLEY, OAT BRAN FLAKES	NFS, GRANOLA
KELLOGGS COCOA KRISPIES CHOCONIL, RICE	KELLOGG'S, RICE KRISPIES	COCOA KRISPIES
KELLOGGS COCOA KRISPIES, RICE	KELLOGG'S, RICE KRISPIES	COCOA KRISPIES
KELLOGGS COCOA KRISPIES, WHEAT	KELLOGG'S, RICE KRISPIES	COCOA KRISPIES
KELLOGGS CRACKLIN OAT BRAN, OAT BRAN	KELLOGG'S, RAISIN BRAN	CRACKLIN' OAT BRAN
KELLOGGS CRISPIX, CORN AND RICE	KELLOGG'S, RAISIN BRAN	CRISPIX
KELLOGGS FROSTED KRISPIES, RICE	KELLOGG'S, RICE KRISPIES	KELLOGG'S, FROSTED RICE KRISPIES
KELLOGGS FRUIT HARVEST, WHOLE WHEAT AND RICE	KELLOGG'S, FRUIT HARVEST CEREAL	
KELLOGGS HONEY CRUNCH CORN FLAKES, CORN	KELLOGG'S, HONEY CRUNCH CORN FLAKES	
KELLOGGS RAISIN BRAN CRUNCH, BRAN	KELLOGG'S, RAISIN BRAN CRUNCH	

Continued—

Table B
Sample search table—continued

PD phrase	Automated FNDDS match	Corrected FNDDS match
NATURES PATH OPTIMUM, ORGANIC WHEAT BRAN	NATURE'S PATH, OPTIMUM	
POST GRAPE NUTS, WHEAT, AND BARLEY	POST, RAISIN BRAN	GRAPE-NUTS
POST SELECTS CRANBERRY ALMOND CRUNCH, MULTIGRAIN	POST, CRANBERRY ALMOND CRUNCH	
POST SHREDDED WHEAT, WHEAT	100%, SHREDDED WHEAT	
POST SUPER GOLDEN CRISP, WHEAT	POST, WAFFLE CRISP	GOLDEN CRISP (FORMERLY CALLED SUPER GOLDEN CRISP)
QUAKER SHREDDED WHEAT, WHEAT	100%, SHREDDED WHEAT	

Note: PD = product dictionary.
Source: Compiled by the authors using data from the IRI product dictionaries and Food and Nutrient Database for Dietary Studies (FNDDS) 2011-12.

Appendix C: The Tradeoff Between Errors and Inclusion

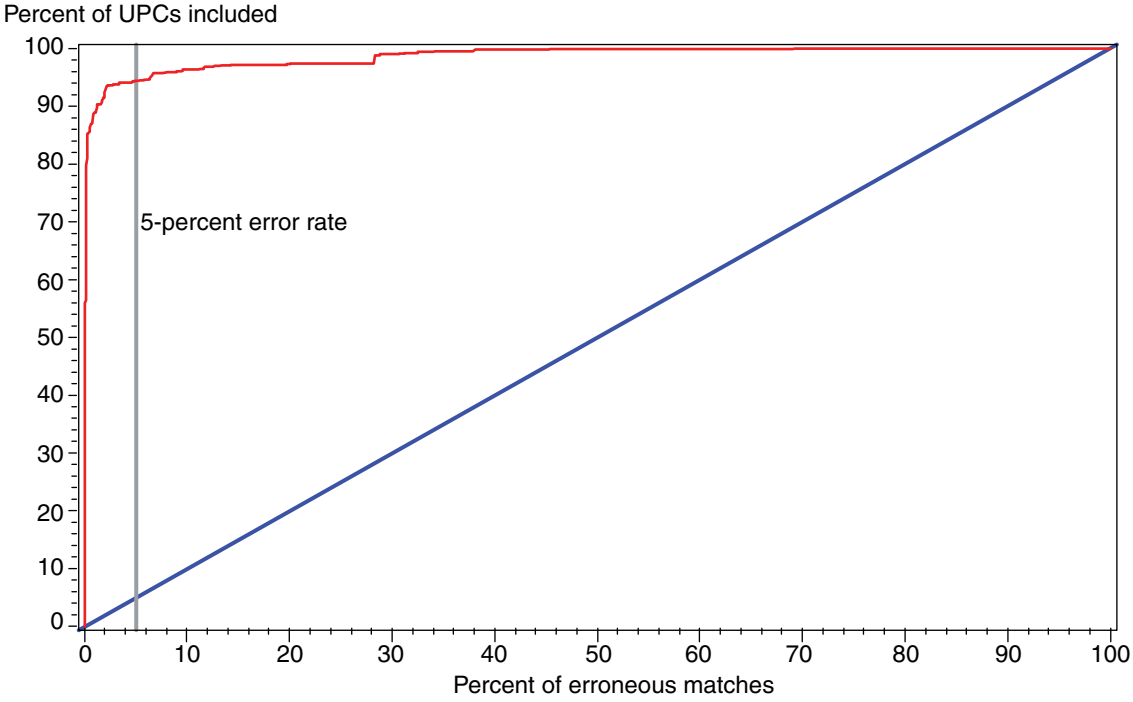
One way to measure the terms of the tradeoff between a low error rate and high inclusion rate is by computing a Receiver Operating Characteristic (ROC) curve (Hajian-Tilaki, 2013) based on a logistic regression model. The ROC curve estimates the terms of the tradeoff between the marginally less accurate prediction of true cases and less accurate prediction of nonmatches.

In the ROC estimates, we regressed the similarity scores (from comparisons of different dimensions of food descriptions from the PD and from FNDDS) on the outcomes of the USDA-established matches from two sets of matches prepared by USDA for other projects: the CNPP Food Prices Database (Carlson et al., 2008) for bread and cold cereal and The Cost of Fruits and Vegetables (Stewart et al., 2016) for fruits and vegetables, which we reviewed and verified prior to use. Within each category, we also included comparisons of almost all incorrect UPC–FNDDS code pairs in the regression.

The area between the ROC curve for cereal (see appendix fig. 1, shown in red) and the 45-degree line (shown in blue) indicates the excellent predictive power of the model. Note that the 45-degree line represents the worst outcome, where including 100 percent of the matches would mean a 100-percent error rate. If we had solved the problem perfectly, the ROC curve would form a triangle with the 45-degree line. The predictive power is near 99 percent overall, including both correctly predicted matches and correctly predicted nonmatches. The curve also demonstrates the tradeoff between specificity (correct links) and sensitivity (fewer omitted links) at the end of the curves. Note that a desired error rate of 5 percent forces us to accept a higher level of omissions than if we had selected a higher error rate. This situation arises pervasively in large-scale linkage problems because a small increase in the ability to predict more correct matches among the very large number of incorrect matches (every pairing of UPC-FNDDS code except the small number of correctly matching pairings) limits the extent to which a linkage method can predict correct matches.

Appendix figure 1

Receiver Operating Characteristic (ROC) curve for the crosswalk



Notes: The ROC curve (red) demonstrates the tradeoff between including more UPCs and erroneous matches in the cereal category. The 45-degree line (blue) represents the worst outcome, where including 100 percent of the matches would mean a 100 percent error rate. The ROC curve compares matches established between the 2013 IRI product dictionaries and the 2011-12 Food and Nutrient Database for Dietary Studies (FNDDS), to verified cereal matches in USDA, Center for Nutrition Policy and Promotion, Food Prices Database. UPC = universal product code. Source: Author estimate using data from the 2011-12 FNDDS, CNPP Food Prices Database, and IRI product dictionaries.
Source: Author estimate using data from the 2011-12 FNDDS, CNPP Food Prices Database, and IRI product

Appendix D: Probability Proportional to Size Sampling and Estimated Error Rate

We used the sampling method probability proportional to size (PPS) to select the review samples to allow items with higher sales volumes within a linking category to be more likely to be selected for review.

In PPS, the probability, π, of item *i* being selected is $\pi_i = nQ_i / \sum_{i=1}^{N} Q_i$, where *n* is the number of food items that will be selected from the linking category, Q_i is the volume sold of food item i (in grams), and *N* is the total number of food items in the linking category. In a PPS sample, we estimate the true error rate, *R*, by dividing the total number of erroneous food items by the number of food items selected for review. To demonstrate that this simple estimate ($\hat{r}$) is a reasonable estimate for the true error rate, *R*, in a linking category, we multiply equation 1.1 by $1 = \pi_i / \pi_i$ and substtute $\pi_i = nQ_i / \sum_{i=1}^{N} Q_i$ into equation 1.1. The estimated error proportion in a PPS sample reduces to

$$\hat{r} = \frac{\sum_{i=1}^{n} (Q_i e_i) / \pi_i}{\sum_{i=1}^{n} Q_i / \pi_i} = \frac{\sum_{i=1}^{n} (Q_i e_i) / (nQ_i / \sum_{i=1}^{N} Q_i)}{\sum_{i=1}^{n} Q_i / (nQ_i / \sum_{i=1}^{N} Q_i)} = \frac{\sum_{1=1}^{n} e_i}{n} \tag{1.2}$$

or the number of errors in the reviewed linking category divided by the number of total items selected.

Since we did not review all possible matches, the estimate, $\hat{r}$, of the error rate changes depending on which food items were selected into the random sample. To estimate the variance for $\hat{r}$ we assumed that *n* is small relative to *N* and that we sampled with replacement to estimate the variance. The estimated variance is thus:

$$\hat{V}(\hat{r}) = \frac{\hat{r}(1-\hat{r})}{n-1} \tag{1.3}$$

where $\hat{r}$ is the estimated error rate parameter from equation 1.2 and *n* is the number of food items in the sample.

Appendix E: List of Acronyms

ARS – USDA Agricultural Research Service

CNPP – USDA Center for Nutrition Policy and Promotion

DGA – *Dietary Guidelines for Americans*

FNDDS – USDA's Food and Nutrient Database for Dietary Studies

FoodAPS – National Household Food Acquisition and Purchase Survey

FPED/FPID – USDA's Food Patterns Equivalents Database/Food Patterns Equivalents Ingredient Database

HEI – Healthy Eating Index

HHS – U.S. Department of Health and Human Services

NHANES – National Health and Nutrition Examination Survey

PD – Product Dictionary

POS PD – IRI Point of Sale Product Dictionary

PPS – probability proportional to size

ROC – Receiver Operating Characteristic Curve

SR – National Nutrient Database for Standard Reference

UPC – Universal Product Code