Methodologies for creating the UNC Food Research Program
US Packaged Food Purchase and Price (PFPP) database
PRELIMINARY

Drafted by: Shu Wen Ng & Izabela Annis
Last updated: December 8th, 2011

1. OVERVIEW

As part of the RWJF funded evaluation of the HWCF marketplace commitment, there is a need to determine the impact of the HWCF commitment on calories purchased by households with children. The simplest approach to evaluate this would be to compare the pre-post commitment calories from HWCF companies' products sold in stores or purchased by household with children. However, even if the HWCF companies change their products and reduce the calories they sell, substitutions to calories purchased from other food companies' (non-HWCF) and retailers' (private label) products may result in no or little change in overall calories purchased. In addition, we know that real income and price changes, shifts in regulations, taxation and other exogenous shifts affect food purchasing patterns over time. These include global and domestic food price changes (around mid-late 2007) and the economic downturn since 2008, which affected both food prices and income differentially. Moreover, there are variation in living costs and market conditions across the US (e.g., Detroit was one of the most economically affected areas), and certain policy (state or local) changes that may have affected access to certain foods should be accounted for. Therefore, both descriptive trend analyses as well as estimation models that take into account prices, income, geographical and other key demographic characteristics of US consumers, as well as any potential policy changes are critical. There are a number of design issues that we will examine and elaborate in the next six months.

Related to the relative role of the HWCF versus non-HWCF companies' products and the private label (PL) products, we want to answer:

1. Did the total calories purchased/sold per year decline between baseline and 2012 and also between baseline and 2015?
2. Which food groups showed the largest and smallest (absolute and relative) caloric changes? Did they represent key sources of empty calories?
3. Which subpopulations experienced the largest and smallest caloric changes?

Our modeling efforts here can only be reflective of purchases from the packaged foods sectors. This poses other critical limits on the impact of these companies as we will evaluate them, and may be beyond our current scope.

To address the above research questions, we need to know the dollar and caloric share of these 16 HWCF food companies compared to non-HWCF companies and compared to private labels (PL) in total and by food groups for all households with children, and also by demographic characteristics (race/ethnicity/income). In addition, we need to have information about the choices or constraints under which food purchase decisions were made, including the variety of prices they were exposed to, and any food related state or local policies. Using the merged Nielsen Homescan and the combined Gladson + Product Launch Analytics nutrition data, we are creating the ***US Packaged Food Purchase and Price (PFPP) Database***. This document provides the methodologies used to create the database, as well as the modeling considerations involved.

The ***US Packaged Food Purchase and Price (PFPP)*** database will contain representative data of the calories purchased and representative prices from UPC foods items (primarily consumer packaged goods) purchased into the household covering the 48 contiguous United States. For each observation (at household-quarter-food group level), there will be information on the demographic characteristics of the household (household composition, income, education and race/ethnicity of female head, county, metro area), total food expenditure within that quarter, total calories bought within that quarter, calories purchased from each food group within that quarter, calories purchased from HWCF, non-HWCF and PL products within each food group, as well as representative prices (per 100grams) for each of the food groups (weighted average price and separate market-quarter representative price for HWCF branded products, non-HWCF branded products, and PL products for each food group) given the market that the household belongs to. We are also planning to merge on additional information about the state or local food related taxes or policies in the future.

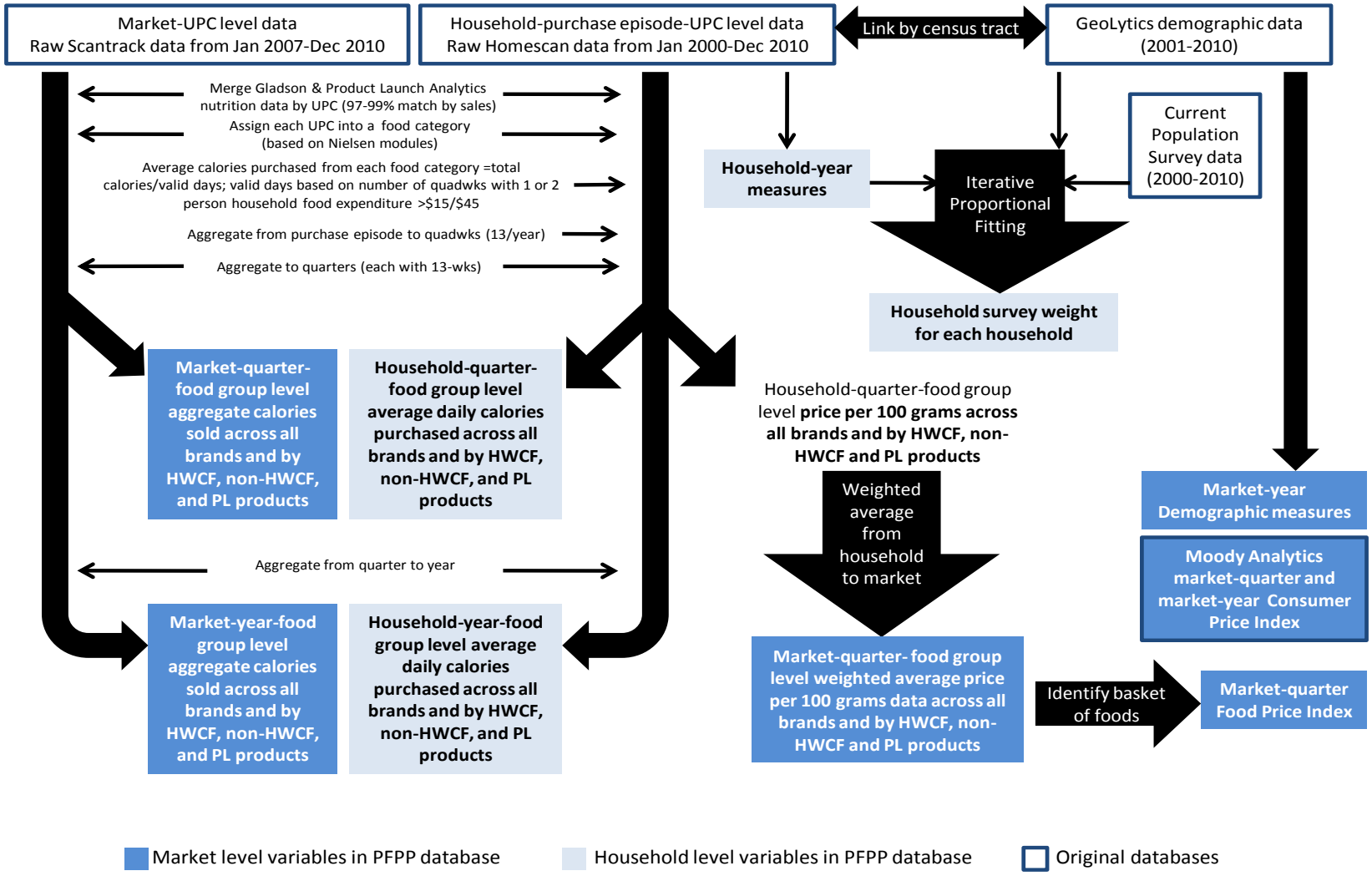
Creation of the PFPP database is complex and requires a number of assumptions and decisions along the way (e.g., the creation of price indices, deflating incomes and expenditures), which are laid out in this document. The steps applied to far are outlined in Figure 1. Details regarding the data sources are discussed in Section 2; how the various data sources are merged is discussed in Section 3; the food and beverage categories are discussed in section 4; and how markets and the calculations of average prices are discussed in Section 5.

Because the creation of the PFPP is work in progress, this version does not provide additional details that will be included in later versions. We plan to include the following in future versions:

- Verification and sensitivity tests conducted
- Other potential data sources to consider in supplementing this database
- Clearer determination as to whether and what leve l of the PFPP data will be made publically available for use depending on discussion with Nielsen and other companies' whose data went into creating this database exactly what can be shared.

UNCFRP
UNC Food Research Program
Robert Wood Johnson Foundation
University of North Carolina at Chapel Hill

Figure 1. Steps taken to create the *US Packaged Food Purchase and Price (PFPP)* database



2. DATA SOURCES

2.1. Food purchase/sales and price data

2.1.1. Nielsen Homescan

The Nielsen Company's Homescan data contain detailed UPC-level information about household food purchases brought into the home. The 2000-2010 data include purchase and demographic information for >60,000 households each year, of which about 35,000 to 65,000 households make up the static sample (defined as having reported purchase for at least once in 10 of the 12 months per calendar year). These households reported all UPC transactions from all outlet channels, including grocery, drug, mass-merchandise, club, supercenter, and convenience stores. Information captured on each shopping occasion include:

- Date of purchase
- Retailer/outlet shopped
- Total food and beverage expenditure on a given shopping trip
- Every UPC purchased along with its
  - Price paid: For purchases made in stores tracked by Nielsen Scantrack, Nielsen assigns the store-level, weighted-mean weekly price for the item to the household's purchase; for non-Scantrack stores, panelists directly report the price paid [1]
  - Number of units purchased
  - Deal used (i.e. store deal, coupon, etc.)
  - UPC characteristics, item description and attributes

Below are the numbers of households in the static sample from Homescan:

- 2000: 34754 HHs (8885 are single member HHs)
- 2001: 34927 HHs (9350 are single member HHs)
- 2002: 39485 HHs (10665 are single member HHs)
- 2003: 39858 HHs (10870 are single member HHs)
- 2004: 39619 HHs (10981 are single member HHs)
- 2005: 50834 HHs (13753 are single member HHs)
- 2006: 62187 HHs (15330 are single member HHs)
- 2007: 63422 HHs (15885 are single member HHs)
- 2008: 61552 HHs (16065 are single member HHs)
- 2009: 60394 HHs (15737 are single member HHs)
- 2010: 61120 HHs (16071 are single member HHs)

Other than sample size growth, the other major change in data collection during the 2000-2010 period was that starting in 2007, Nielsen stopped collecting details about random weights by actual food items (e.g., apples, banana, broccoli, tomatoes, beef, pork, fish, shrimp, etc.). Instead, Nielsen groups all fruits and vegetables in one category, all deli meats and cheeses in one category, and all meats (poultry, beef, pork, seafood) in one category.

Nielsen calculates household-level weights for households in the static sample to project a demographically balanced panel to match the US population as closely as possible at the metropolitan

market, regional, and national level using census demographic information for each geographic area [2]. However, there are concerns about the Nielsen household-level projection factors and whether these are truly representative of the US population since Nielsen's sampling frame and response rate are unknown outside of Nielsen. Therefore, the PFNPP will create its own weight for each household by matching measures in Homescan with measures available in the data sources discussed in Section 2.2. We will compare the PFNPP household weights and Nielsen's household projection factors to see how well they correlate (details and results in Section 6.2)

Household level information in Nielsen Homescan include:

- Household income
- Household composition (size, number of children, age and gender of each household member)
- Race of head of household (male if available, otherwise female)
- Ethnicity of head of household (male if available, otherwise female)
- Highest educational attainment of male head
- Highest educational attainment of female head
- Occupation of head of household
- Geographic identifiers (market, census division, census region)

Other limitations and concerns related to the Nielsen Homescan data in capturing purchases of packaged foods include:

- Selection bias in response rates, participation and attrition. For example, households with older heads (especially retirees) are more likely to participate and less likely to attrit. Similarly, there might be less than proportional representiveness of households with children (creating weights to adjust for population distributions will not correct for the selection bias)
- Misreporting by household respondents
- No information about away-from-home food and beverage purchases that were never brought into the house by all members of the household
- Information on random weight items is very limited and aggregated since 2007 (but PFNPP only focuses on packaged foods)
- Household level information is only updated annually

2.1.2. Nielsen Scantrack

Store-based scanner data, such as Nielsen Scantrack, provide records of weekly dollar sales and units sold of all UPC transactions at participating grocery, drug, mass merchandiser stores and convenience stores.

Like Homescan, there are also concerns about the representativeness of the Scantrack data in measuring sales of packaged foods in the US for a number of reasons. First, there are a large proportion of stores are missing in Scantrack – warehouse club stores (e.g., Costco, BJ's), large grocery stores chains such as Walmart, Whole Foods, Aldi, Trader Joe's are missed. Therefore, any results derived purely from the Scantrack data is only representative of the stores included in their sample. Second, it does not have any information about sales of packaged food and beverages from vending machines, restaurants, food trucks, mom and pop stores, ethnic markets, specialty markets, etc. Third, as shown in Table 2.1 only five markets have information from the four channels Scantrack collects its data from, which would make it challenging to make projections to the national level.

Table 2.1 Markets and channels covered by Nielsen Scantrack

Markets (52 in total)	Food/Grocery >$2 mm	Drug >$1mm	Mass	Convenience
5 markets: Atlanta, Chicago, Los Angeles, Philadelphia, San Francisco	X	X	X	X
4 markets: Boston, Detroit, Houston, New York	X	X		X
2 markets: Pittsburgh, Washington DC	X	X	X	
21 markets: Birmingham, Cincinnati, Cleveland, Dallas, Denver, Little Rock, Louisville, Miami, Minneapolis, Nashville, New Orleans/Mobile, Oklahoma City/Tulsa, Orlando, Phoenix, Portland, Raleigh/Durham, Richmond, San Antonio, Seattle, St. Louis, Tampa	X			X
20 markets: Albany, Baltimore, Buffalo/Rochester, Charlotte, Columbus, Des Moines, Grand Rapids, Hartford/New Haven, Indianapolis, Jacksonville, Kansas City, Las Vegas, Memphis, Milwaukee, Omaha, Sacramento, Salt Lake City/ Boise, San Diego, Syracuse, West Texas	X			
According to Nielsen, % of total US store universe	81.9%	37.6%	16.3%	
According to Nielsen, # of stores in sample	3,025	700	370	7,500
According to Nielsen, # of stores represented	31,785	27,751	3,867	144,648
Nielsen collection method	Scan	Scan	Scan	50% scan; 50% audit

Other limitations and concerns related to the Nielsen Scantrack data include:

- Store brand/ Private Label (PL) items' UPCs are masked (e.g., regular Kroger Big K cola and regular Safeway Go2 cola will be all under one 'fake' UPC to denote PL regular colas).
- The price estimates derived from Scantrack and applied to Homescan (already done by Nielsen) may be significantly upward biased because the Scantrack data do not include price data from discount supercenter or warehouse club stores such as Walmart and Costco, as well as prices of each of the PL items, which current estimates suggest comprise over 30 percent of consumer food-at-home expenditures.

For the purposes of creating the PFNPP database we will not be using the Scantrack data directly for the creation of prices, but will be considering it as a source of additional information for determining how much of the sales data is reflected in calories purchases and therefore provide a minimum degree to which calculations on calories purchased as measured in Homescan should be adjusted. For example, we can compare the calories purchased by each market at each quarter derived from Homescan to the calories sold at each market at each quarter for the five markets that has information from all four channels in the Scantrack data. Similarly, it is possible to use the Homescan data to supplement the gaps in the Scantrack data related to missing major stores. It is important to note that both Homescan and Scantrack will be underestimates of the absolute calories and dollar purchase or sales of the US population. Sales data directly from the HWCF companies will allow us to at least determine for these companies the total sales from all venues to try to adjust for all others.

2.2. Current Population Survey (CPS)

The CPS is the primary source of information on the labor force characteristics of the US population. The sample is scientifically selected to represent the civilian non-institutional population and Armed Forces personnel living off post or with their families on post. Respondents are interviewed to obtain information about the employment status of each member of the household 15 years of age and older. However, published data focus on those > 16 years old. The sample provides estimates for the nation as a whole and serves as part of model-based estimates for individual states and other geographic areas.

Estimates obtained from the CPS include characteristics of households and families, employment, unemployment, earnings, hours of work, and other indicators. They are available by a variety of demographic characteristics including age, sex, race, marital status, and educational attainment. They are also available by occupation, industry, and class of worker. Supplemental questions to produce estimates on a variety of topics including school enrollment, income, previous work experience, health, employee benefits, and work schedules are also often added to the regular CPS questionnaire. More information can be found at: http://www.census.gov/prod/2006pubs/tp-66.pdf

2.3. Consumer Price Indices

The Consumer Price Index (CPI) is a measure of the average change over time in the prices paid by consumers for a market basket of consumer goods and services. The CPI can be used as an economic indicator, as a deflator of other economic series, and as a means of adjusting dollar values (see: http://www.bls.gov/cpi/cpiadd.htm#2_1).

2.3.1. Moody's Analytics

Moody's Analytics (MA) estimates monthly seasonally adjusted and unadjusted consumer price indices (CPI) for the metropolitan areas in the US. These estimates are derived from the BLS monthly national, regional and area consumer prices and MA's estimates of metropolitan cost of living and median household income. An initial non-seasonally adjusted estimate of metropolitan CPIs is created by relating MA's cost of living indices to national CPI. This series is then back-casted and extended using census region population sized CPIs adjusted by the ratio of population sized regional median household income by the metropolitan median household income. This series is then indexed to average 100 between 1982 and 1984.

The BLS publishes indexes for 27 local areas. These indexes are byproducts of the national CPI program. MA maps these CPIs to OMB defined metropolitan areas and replace our initial estimates with these indexes for 27 areas. Next, all the estimates are seasonally adjusted using the US Census Bureau's X-12 Arima program. Finally, seasonally adjusted indices are created for the eleven metropolitan areas with metropolitan divisions by weighting the CPI for the 29 metropolitan divisions by population. More information can be found from the U.S. Metropolitan Areas Forecast Database: http://www.economy.com/home/products/brochures/Databases-Metro-Forecast.pdf

2.4. USDA Standard Reference

The USDA National Nutrient Database for Standard Reference (SR) is the major source of food composition data in the United States. It provides the foundation for most food composition databases

in the public and private sectors. As information is updated, new versions of the database are released. We applied the congruent years of the SR to each of the Nielsen datasets when needed. The most recent version, Release 24 (SR24) release in September 2011, contains data on 7,906 food items and up to 146 food components. These include data for raw, processed, and prepared foods including raw and cooked animal (e.g., recently updated cooked beef cuts) and vegetable foods, grain and baked products, selected brand name fast food items, including burgers and pizzas, soups, breakfast cereals, and candies, etc. Data have been compiled from published and unpublished sources. Published sources include the scientific literature. Unpublished data include those obtained from the food industry, other government agencies, and research conducted under contracts initiated by USDA's Agricultural Research Service (ARS). Values in the database may be based on the results of laboratory analyses or calculated by using appropriate algorithms, factors, or recipes, as indicated by the source code in the Nutrient Data file. Not every food item contains a complete nutrient profile [3]. The key limitations of the SR for the use on packaged foods are that the range of foods included are limited, many of the nutrition information is for foods in raw or unprepared states, the SR only started including saturated fat starting in SR 10 (release 1993), and started including total sugars only since SR 18 (released 2005).

2.5. Nutrition Label databases

Nutrition Facts Panel (NFP) labeling is required for most prepared foods, such as breads, cereals, canned and frozen foods, snacks, desserts, and drinks. Per the Food and Drug Administration (FDA) requirements, the NFP need to include serving measurement, total calories, calories from fat, total fat, saturated fat, sugars, total carbohydrate, protein, dietary fiber, sodium, cholesterol, vitamin A, vitamin C, calcium and iron.

However, NFP data sources also have limitations. First, while NFP labeling is required for most prepared foods, nutrition labeling for raw produce (fruits and vegetables) and fish is voluntary. In addition, delicatessen type foods, bakery products and confections sold directly to consumers from preparation locale, and self-service bulk foods do not have NFP labeling requirements at all. Thus, the USDA SR may be especially useful for such foods, especially since they generally will not change significantly every year or two in their nutrient content (especially for raw fruits and vegetables).

Second, the information on the NFP is not as precise as what can be found from the USDA SR data sources due to reporting and rounding rules to allow for space limitations and ease of communicating to consumers. For example, foods with less than 5 calories or 5 gram of fat per serving meet the definition of "calorie free" or 0 grams of fat respectively, foods with 50 calories or less per serving are rounded to the nearest 5-calorie increment, and foods above 50 calories per serving are round to the nearest 10-calorie increment [4]. This allows manufacturers to use serving size measurements that can result in their entire package be listed has having 0 grams of fat, when it is not necessarily the case. Additional details about the NFP or food labeling regulations that should be noted include the fact that only the nutritional properties of the product as packaged is required, although nutritional information may be voluntarily presented "as prepared", and that there is a 20% 'measurement allowance' between what is shown on the NFP and what is found during an enforcement analysis. Complete details on the NFP labeling regulations can be found in Code of Federal Regulations Title 21 [4].

2.5.1. Gladson Nutrition Database (Gladson)

The Gladson Nutrition Database contains close to 175,000 records of comprehensive packaging and labeling information. These include all the nutrition facts panel information, full ingredient

listings, serving size, servings per package, warning, nutrition or health claims on the packaging, date of last update, manufacturer and brand by UPC. We are only using the NFP data for the package as it (not as prepared) in order to prevent double counting of using additional ingredients in food preparation. These will be merged by UPC and year with the appropriate years of Nielsen Homescan, Scantrack and HWCF/GES sales data. Based on our experience working with this database, there are at least two concerns:

- New nutritional information for the same UPC overwrites over prior information for that UPC. Therefore, there will be measurement error. For example, for UPCs that were reformulated in 2007, we can only use the 2007 Gladson data for 2005 and 2006 Nielsen data, which may not be accurate measures of the nutritional content of those UPCs in 2005 and 2006.
- While the Gladson data indicates when the information for a particular UPC was last updated, the degree and frequency of when and how comprehensively the commercial databases are updated is unclear and appears inconsistent.

2.5.2. Mintel Global New Product Database (GNPD)

Mintel Global New Product Database (GNPD) monitors the introduction of new products globally at the UPC level. It contains over 96,000 records with usable nutrition information (reported serving size, package weight and total calories per serving). We will continue using pulling new usable records for each calendar year.

2.5.3. Datamonitor Product Launch Analytics (PLA)

Datamonitor's Product Launch Analytics (PLA) monitors the introduction of new products globally at the UPC level. Since 2009, it started collecting some nutrition information for food and beverage products. For the period of January, 2009 – May 2010, we identified over 7,000 records with usable nutrition information (reported serving size, package weight and total calories per serving).

3. MERGING FOOD PURCHASE AND PRICE DATA WITH THE NUTRITION DATABASES

In order to derive the nutritional content purchased or sold, it is necessary to link the food purchase and price data with the nutrition databases.

This is a multi-step process outlined below:

1. Direct "by UPC" merge of Gladson, Mintel and PLA to Homescan & Scantrack based on date of last update and Homescan/Scantrack data year.

2. For every UPC that is unmatched after Step 1, we find a "sister" UPC (that was already linked in Step 1) that has the same brand, module and product description. We use sister's UPC nutritional facts adjusting for size differences.

3. For every UPC that is unmatched after Step 2, we find a sister UPC that has the same brand, module and "short" product description, where "short "product description is the product description stripped of multipack information (if any exists).

4. For every UPC that is unmatched after Step 3, we find a sister UPC that has the same brand, module and "short" product description, where "short " product description is the product description stripped of some nutritionally irrelevant information (i.e. for instance, one can safely omit "CN" – can and "NBT" – non-refillable bottle from product description in the soft drinks module and match UPCs on the remaining information in the product description).

5. For every UPC that is unmatched after Step 4, for the products that are in a relatively simple/"raw" state (i.e. dry barley), we find an already matched sister UPC in the same module. Certain categories (i.e. eggs, milk, plain, unflavored cottage cheese) also require taking into consideration certain nutritionally relevant product attributes. For instance, all unmatched jumbo eggs UPCs where matched to already matched jumbo egg product (i.e. module and size of egg are taken into consideration). However, any matched product in "BAKING POWDER" module will be linked to already matched UPC in the "BAKING POWDER" module (i.e. only module is taken into account). First attempt is to match products within the same brand (step coded as 5.1). If no matches are found then the search for matching product is expanded beyond brand (coded as 5.2).

6. Many products that have either zero or close to zero caloric value do not have nutrition labels. We identified several categories that did not have matching nutrition data and we are assuming that their total calories equal to 0 (i.e. ground coffee).

7. For every UPC that is unmatched after Step 6, we find a sister UPC that has the same module and one or more of the nutritionally relevant characteristics. For instance, for MILK-SHELF STABLE we chose flavor and type. While for NUTS-BAGS, we chose salt attribute and search UPC description for nuts types (i.e. walnuts, peanuts, etc.) First attempt is to match products taking into consideration brand (step coded as 7.1). If no matches are found then the search for matching product is expanded beyond brand (coded as 7.2).

8. All the UPCs in the DEPT-ALCOHOL, that are unmatched after step 7, are linked to available USDA SR data. For instance all hard liquor, such as "BOURBON-BLENDED", "RUM", "VODKA", etc. with proof

of 100 are linked to SR num = 14533 (alcoholic beverage, distilled, all, 100 proof).
Broad assumptions are made about proof ranges as nutrition data are extremely limited.

Nielsen to Nutrition Facts Panel match by Dollar Volume sales

Year	Homescan	Scantrack
2000	96.9%	No data
2001	97.0%	
2002	97.0%	
2003	97.0%	
2004	96.9%	
2005	97.4%	
2006	96.7%	
2007	97.0%	96.8%
2008	97.3%	96.9%
2009	98.1%	98.0%
2010	98.2%	98.0%

4. FOOD AND BEVERAGE CATEGORIES

We classified all UPCs into one of 62 mutually exclusive food and beverage categories based on information provided by Nielsen on department (dairy, dry grocery, frozen, meat, deli, fresh produce, or alcohol), product group (64 unique product groups) and module (over 700 unique modules).

Table 4.1 Food Categories applied to Scantrack and Homescan

Dairy products excluding milk	**Sweets & snacks**
Cheese	RTE cereal bars and toaster pastries
Yogurt	RTE grain-based desserts
Frozen/refrigerated dairy-based toppings/condiments	Cookies
Shelf-stable creamers, evaporated or condensed milks	Crackers
	Shelf-stable snacks
Meat, Poultry, fish & mixtures	Spreads and dips
Refrigerated seafood	Candy & gum
Frozen seafood	Frozen/refrigerated pudding and ice cream
Frozen & fresh meat & poultry	Shelf-stable pudding and gelatin
Canned seafood & poultry	Shelf-stable dessert toppings
Canned processed meats	Sweeteners
Refrigerated processed meats	Nut and fruit spreads
Other protein sources	**Other**
Fresh eggs	Baby food
Nuts & seeds	Baking supplies
	Spices, seasoning, & extracts
Grain products, no RTE desserts	
Cereals (requires cooking)	**Mixed dishes and soups**
RTE cereals and granola	Frozen entrees
Boxed, dry pasta & rice	RTE, prepared dishes
Fresh/frozen pasta	Canned mixed dishes
Boxed pasta & rice dinners	Shelf-stable soups & stews
Shelf-stable Mexican-style products	
Dry baking mixes	**Beverages**
Flours	Fresh plain milk
RTE breads	Refrigerated sweetened dairy drinks
Frozen baked goods	Shelf-stable milks, milk substitutes and milk-based powders
Refrigerated/frozen dough products	Shelf-stable fruit and vegetable drinks and juice
Frozen/refrigerated breakfast products	Frozen fruit drinks and juice
RTE sandwiches	Beverage powder and concentrates
Frozen/refrigerated pizza and appetizers	Carbonated soft drinks
	Tea (bags, loose, RTD)
Fruits & vegetables	Coffee (grounds, beans, RTD)
Fresh & frozen fruit	Water and ice
Canned/dried fruit	Alcohol
Canned/dried vegetables & legumes	
Fresh & frozen vegetables	
Fats, oils, sauces & condiments	
Fats and oils	
Condiments, dressing & sauces	

5. DEFINING MARKETS, CALCULATING AVERAGE PRICES AND CREATING FOOD PRICE INDICES

5.1. Defining Markets

Nielsen constructs the Homescan data by including households from both metropolitan (metro) and non-metropolitan (remaining) areas in order to create a national sample of households. Nielsen metro households are defined as living in one of 52 large metropolitan areas, while remaining households are those residing outside of one of those 52 areas, which come from across the 9 census divisions (see Table 5.1). We are able to retain these markets in the PFPP since our focus is on packaged foods and beverages (unlike the ERS's QFAHPD, which needed to collapse a number of the markets due to sample size issues with the fresh foods).

Table 5.1 Markets included

52 Metro markets		Remaining markets (non-metro)
Boston	San Diego	Rem Greenville
Chicago	St. Louis	Rem Atlanta
Houston	Tampa	Rem Boston
Indianapolis	Baltimore	Rem Charlotte
Jacksonville	Birmingham	Rem Denver
Kansas City	Buffalo-Rochester	Rem Detroit
Los Angeles	Hartford-New Haven	Rem Indianapolis
Surburban New York	Little Rock	Rem Jacksonville
Urban New York	Memphis	Rem Kansas City
Exurban New York	New Orleans-Mobile	Rem Memphis-Little Rock
Orlando	Oklahoma City-Tulsa	Rem Milwaukee
San Francisco	Phoenix	Rem Minneapolis
Seattle	Raleigh-Durham	Rem New Orleans-Mobile
Atlanta	Salt Lake City	Rem North California
Cincinnati	Columbus	Rem Oklahoma City-Tulsa
Cleveland	Washington, Dc	Rem Omaha
Dallas	Albany	Rem Philadelphia
Denver	Charlotte	Rem Pittsburgh
Detroit	Des Moines	Rem Richmond-Norfolk
Miami	Grand Rapids	Rem San Antonio-Albuquerque
Milwaukee	Louisville	Rem Seattle-Portalnd
Minneapolis	Omaha	Rem St. Louis
Nashville	Richmond	Rem Los Angeles-Collar
Philadelphia	Sacramento	Las Vegas
Pittsburgh	San Antonio	
Portland, Or	Syracuse	

5.2. Calculating average prices

We are interested in creating four types of prices for each food group at each market and quarter:

- Average HWCF company brand price for each food group for each market and quarter
- Average non-HWCF company brand price for each food group for each market and quarter
- Average store brand/private label (PL) price for each food group for each market and quarter
- Average price for all items from each food group for each market and quarter

The Nielsen Homescan data provide detailed information about each food purchase, including number of units or packages, total weight, and total amount paid at the UPC level. Using this information, we calculate the price per 100 grams (unit value) for each purchase of each UPC food item. For dry weights, we use a conversion of 28.35 grams per ounce, and a conversion factor of 29.57 grams per ounce for liquids. In some cases, however, only the counts purchased (e.g., ears of corn) is reported. In these cases, we used the USDA National Nutrient Database for Standard Reference (Release 23) to convert the unit counts to weight, assuming the food was medium-sized (if there are multiple sizes in the database). Although it was possible to convert most unit counts to gram weights using this approach, not all purchases reported only as counts were convertible. Those food items that were not converted were excluded from the price calculations. We did not reduce the weight of foods purchased by the amount of the food that is inedible since our database is constructed for prices of foods as purchased, not as consumed.

5.2.1. Average company brand's price for each food group-market-quarter

A purchase event (p) is uniquely identified in Homescan data as household (h) reporting a purchase of UPC (i) on date (d) in store (s). To create the quarterly (q) average company (c) price per 100 grams for each food group (g) in market (m) we aggregate the total expenditure on all UPCs from food group (g), company (c) for all households (h) residing in market (m) during a particular quarter (q), weighting by household (h) specific weight.

$$p_{c,m,q,g} = \frac{\sum_{j=1}^{n} HHweight_h * p_{j,h}}{\sum_{j=1}^{n} HHweight_h * I_{j,h}}$$

Where household (h) ε market (m), purchase (j) ε company (c), food group (g), quarter (q).

5.2.2. Average price for each food group-market-quarter

We applied the same strategy mentioned above but did not distinguish by company brand. This average price for each food group-market-quarter is essentially a weighted average of the three average company brand's price.

5.3. Creating a food price index

Besides a Consumer Price Index (CPI), it is also important to have a Food Price Index so that we can take into account the relative prices of different kinds of food.

In creating the Food price Index, we include food and beverage groups for which there are at least 10 households or 1% of all static households in a market (whichever is higher) that purchase items from that food groups across all the markets and quarters. Imposing this sample size restriction results in exclusion of 5 food groups (refrigerated seafood, fresh/frozen pasta, baby food, shelf stable milks, and alcohol).

Like the BLS, the average per 100 gram price of these food and beverage groups *(k)* were then weighted based on the proportion of food expenditure (FG weight) on each of these food groups at the market-quarter level based on the formulae:

$$FPI_{m,q} = \frac{\sum_{k=1}^{n} FGweight_{m,q,k} * p_{m,q,k}}{\sum_{k=1}^{n} FGweight_{m,q,k}}$$

We use Los Angeles first available quarter of 2000 Q1 and use it as a base FPI (i.e., 2000 Q1, Los Angeles = 100). The FPI for all other market- quarters are relative to this base.

REFERENCES

1. Einav L, Leibtag E, Nevo A. On the Accuracy of Nielsen Homescan Data. Washington, DC: USDA; 2008.
2. Zhen C, Taylor JL, Muth MK, Leibtag E. Understanding Differences in Self-Reported Expenditures between Household Scanner Data and Diary Survey Data: A Comparison of Homescan and Consumer Expenditure Survey. Rev Ag Econ. 2009; **31**(3): 470-92.
3. Agricultural Research Service. USDA National Nutrient Database for Standard Reference. In: US Department of Agriculture, editor. Release 23. Beltsville, MD: USDA; 2010.
4. National Archives and Records Administration. 21 CFR 101.9. Washington, DC; 1993.