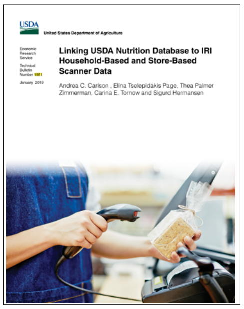
United States Department of Agriculture

A report summary from the Economic Research Service March 2019



United States Department of Agriculture

Economic Research Service

Technical Bulletin Number 1951

January 2019

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

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

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.

ERS is a primary source of economic research and analysis from the U.S. Department of Agriculture, providing timely information on economic and policy issues related to agriculture, food, the environment, and rural America.

- 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.