

# Assessment of Commercial Store and Household Scanner Data: Methods, Content, and Cautions

Presented at FCSM/WSS Workshop on Transparent Reporting on the Quality of Integrated Data, Washington, DC, December 1, 2017

Mary K. Muth

Director, Food, Nutrition, & Obesity Policy Research Program



United States Department of Agriculture

Economic  
Research  
Service

Technical  
Bulletin 1942

April 2016

## Understanding IRI Household-Based and Store-Based Scanner Data

Mary K. Muth, Megan Sweitzer, Derick Brown,  
Kristen Capogrossi, Shawn Karns, David Levin,  
Abigail Okrent, Peter Siegel, and Chen Zhen



United States Department of Agriculture

Economic  
Research  
Service

Technical  
Bulletin  
Number 1946

September 2017

## Food-at-Home Expenditures: Comparing Commercial Household Scanner Data From IRI and Government Survey Data

Megan Sweitzer, Derick Brown, Shawn Karns,  
Mary K. Muth, Peter Siegel, and Chen Zhen



# Acknowledgments and Disclaimer

- Most of the content for this presentation was developed under Agreement 58-5000-3-0069 with the U.S. Department of Agriculture (USDA), Economic Research Service (ERS) and under data use agreements with IRI to access the InfoScan and Consumer Network data.
- Collaborators on this work include Megan Sweitzer (ERS), Abigail Okrent (ERS), David Levin (ERS), Shawn Karns (RTI), Peter Siegel (RTI), Derick Brown (RTI), and Chen Zhen (UGA).
- Any opinions, findings, conclusions, or recommendations expressed in this presentation are not attributable to USDA, ERS, or IRI.

# Introduction

- Primary types of scanner data available from commercial suppliers (specifically, IRI and Nielsen in the U.S.)
  - Store-based
  - Household-based
- *Advantages of scanner data*
  - Provide high frequency product prices and purchase quantities at the store-keeping unit (SKU) level
    - By Universal Product Code (UPC) or Price Lookup Code (PLU)
    - By individual household, individual store, or geographic area
- *Considerations in using scanner data*
  - Cost of purchasing or obtaining license to use the data
  - Limited availability of documentation on sampling, data collection, and weighting methods
  - Representativeness depending on particular application
  - Potential restrictions on release of analysis results

# Introduction (continued)

- Examples of current government uses
  - Construct prices for ERS Quarterly Food at Home Price Database
  - Calculate cost of the WIC food package
  - Calculate cost of the Thrifty Food Plan, which is the basis for the SNAP allotment formula (updated using CPI)
- Importance of understanding the properties of the data
  - Sample selection methods
  - Data collection and processing methods
  - Weighting methods
  - Comparisons to other data sources

# IRI InfoScan Store Scanner Data: Contents

- Data obtained from transactions data provided by retailers to IRI
  - Includes IRI “census” stores that have agreed to provide sales for all stores
    - Excludes “sampled” stores that IRI randomly selects from the remainder
  - Includes private label (store brand) sales from selected retailers
    - A few retailers only release data at the brand/category level, which means package size information is not available
  - Some retailers release individual store data while others aggregate to retailer marketing area (RMA)
- Data obtained by ERS represent an unprojected (unweighted) subset of the total IRI store data
- Dataset components:
  - Week
  - Store ID or geography key (RMA-level data)
  - UPC code (indicating package size)
  - Quantity
  - Total value of purchase
  - Can be linked to store and product information

# InfoScan Store Scanner Data: Data Collection Procedures

- IRI receives daily sales data from retailers including products with UPCs and random-weight products
  - Retailers aggregate individual transactions to the UPC or product level
  - IRI aggregates to a weekly level and conducts quality control checks
- Note about random-weight and uniform-weight perishable products (e.g., fresh produce, meat, deli, bakery)
  - Some products are scanned
    - Products with UPC codes (uniform-weight)
    - Products that are pre-weighed and labeled at the store
  - Some products are weighed and product codes are entered by the cashier
    - Products with price lookup codes (PLUs)
- Most retailers report total units sold and total dollars
  - Total dollars are net of loyalty card discounts
  - Can calculate unit prices (e.g. price per ounce) by dividing weighted-average price by number of units in the package

# InfoScan Store Scanner Data: Store Counts

## Numbers of Stores Represented, 2012

	UPC			Random weight		
	Store-level	RMA-level	Total	Store-level	RMA-level	Total
Convenience	9,613	0	9,613	0	0	0
Defense	515	10	525	0	0	0
Dollar	8,237	0	8,237	1,282	0	1,282
Drug	12,497	7,358	19,855	12,176	7,341	19,517
Grocery	7,100	5,743	12,841	6,720	5,743	12,463
Liquor	341	464	805	0	0	0
Mass/ club	3,140	4,521	7,661	1,786	4,485	6,271
<b>Total</b>	<b>41,443</b>	<b>18,096</b>	<b>59,537</b>	<b>21,964</b>	<b>17,569</b>	<b>39,533</b>

# InfoScan Store Scanner Data: Comparisons

## InfoScan Relative to Census Bureau Data, 2012

	Percentage of Stores			Percentage of Sales		
	Store-level	RMA-level	Total	Store-level	RMA-level	Total
Convenience	36%	0%	36%	35%	NA	35%
Dollar	23%	0%	23%	19%	NA	19%
Drug	29%	17%	46%	69%	50%	119%
Grocery	25%	20%	46%	25%	25%	50%
Liquor	1%	1%	2%	2%	4%	7%
Mass/club	61%	88%	150%	9%	70%	79%
<b>Total</b>	<b>28%</b>	<b>12%</b>	<b>41%</b>	<b>22%</b>	<b>34%</b>	<b>55%</b>

*Census Bureau estimates are from the 2012 Economic Census, Industry Series.*



# InfoScan Store Scanner Data: Considerations

- Stores represented in the data
  - Data collection process is not designed to capture sales from smaller, independent stores
- Private-label product data
  - Not provided by all retailers
  - Aggregation of data by some retailers prevents calculation of unit prices
- Random-weight data (e.g., produce, meat, deli, bakery)
  - Only available for some stores
  - Product information is limited
  - Must determine if units are weights or counts
- Projection factors (or weights)
  - Not provided with ERS data; therefore unable to calculate national estimates
  - RTI has a contract to develop weights for use by ERS

# IRI Consumer Network Household Scanner Data: Contents

- Data obtained from the National Consumer Panel (joint venture between IRI and Nielsen)
  - Households are recruited online and complete demographic survey
  - Households are randomly selected to meet quotas by demographic category
  - Household record purchases using an in-home scanner or mobile app
- Data are weighted using a raking (IPF) procedure
- Dataset components:
  - Purchase date
  - Household ID
  - Store ID
  - UPC code
  - Quantity
  - Price (and use of coupons or deals)
  - Projection factor
  - Can be linked to store, household, and product information

# CN Household Data: Household Counts (2012)

- All households in the panel record UPC products and a portion also records random weight products
- Households are included in the annual “static” panel if they meet requirements for
  - Minimum frequency of reporting
  - Minimum average spending level for household size
- Projection factors are calculated for the static panel

Dataset	No. of Households			No. of Transaction Records		
	Static	Total	Static %	Static	Total	Static %
Consumer Network	62,517	126,040	50%	58.8 MM	72.1 MM	82%
Random Weight	33,852	78,992	43%	5.0 MM	6.4 MM	78%

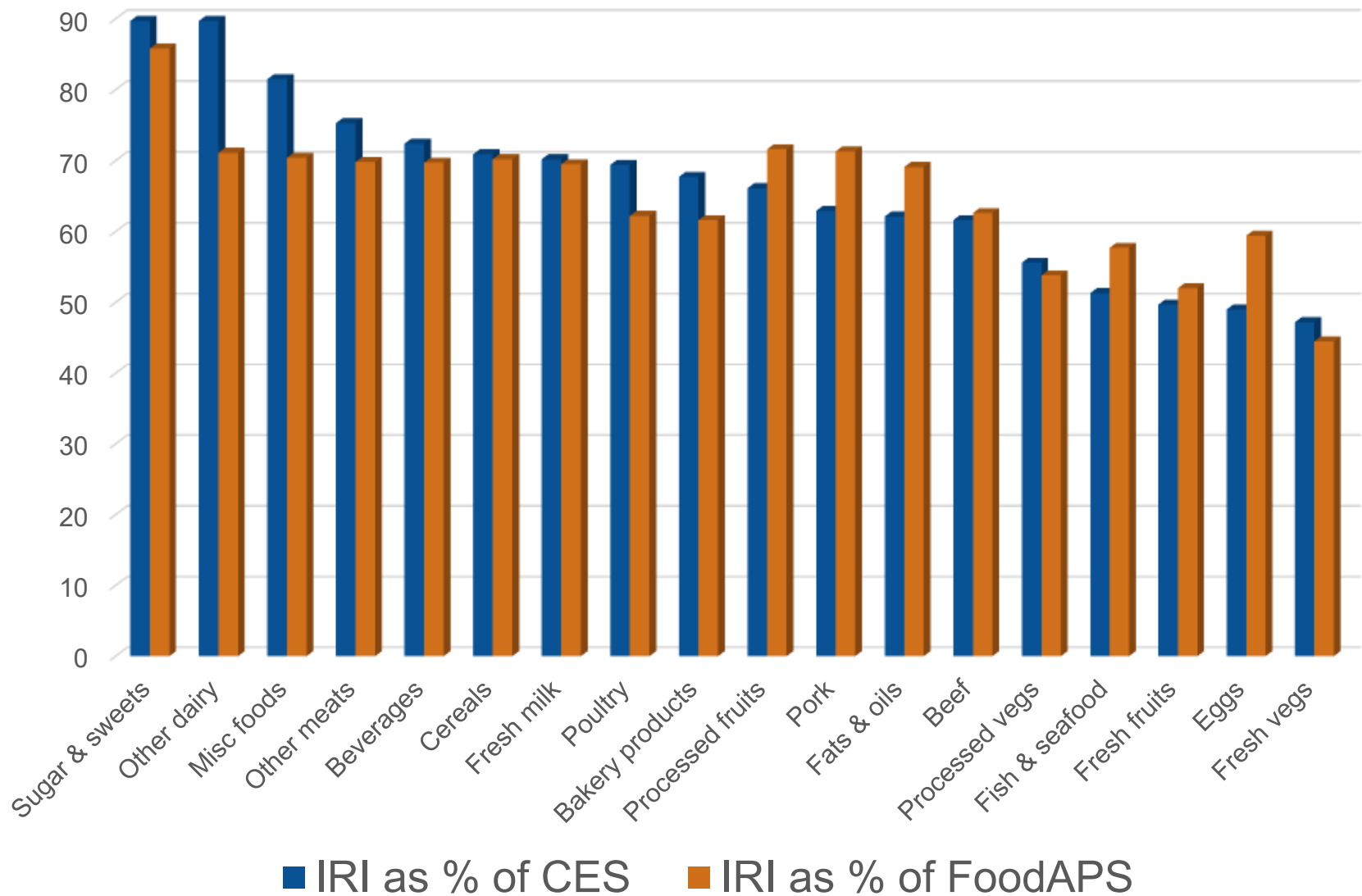
# CN Household Data: Data Collection Procedures

- Purchase recording by households
  - Indicate store where purchased
  - Packaged products—scan UPC; indicate if product on sale or received a deal
  - Random weight products—select from list of products or scan code on reference card and enter total amount paid (no quantities recorded)
  
- IRI price assignment
  - Assigns average price for store chain and market area using store scanner data
    - If not available, assigns average price for store type and market area
  - If no store scanner data, household enters price
  - Last resort, assign “dictionary” price

# CN Household Data: Weighting

- IRI calculates projection factors using Iterative Proportional Fitting
  - Separate weights for entire static panel and static random weight panel
  - Demographic targets are based on Census demographic data (obtained through PopStats™)
    - Household size, age of household head, household income, ethnicity, race, presence of children, county size
- Projection factors are dynamic
  - Households appearing in the data across multiple years have new projection factors each year

# CN Household Data: Comparison of Average Weekly Household Expenditures to Other Sources, 2012



# CN Household Data: Considerations

- Households that participate are likely different from the general population
  - Intensive data collection process
  - More aware consumers
- Some types of households are less like to meet qualifications for inclusion in static panel
  - Younger (under age 35) households
  - Lower income households
  - Black and Hispanic households
  - Households with children
- Prices are typically not exact prices paid by the household
- Data are weighted based on demographics, not shipment or expenditure totals

# Conclusions

- Data are collected for commercial purposes
  - Not necessarily designed for research purposes
- Goals of the data vendors are to:
  - Adhere to agreements with stores regarding level of disclosure
  - Ensure confidentiality of household participants
  - Protect their competitive information
- In using the data, it is important to understand the data collection and processing procedures and assess implications for results of analyses based on:
  - Characteristics of stores and households that participate
  - How quantities, prices, or expenditures are recorded
  - How the weights are constructed (if available)
- But no other comparable data source provides the same level of granularity and detail needed for many types of analyses



**Mary K. Muth, PhD**

919.541.7289

[muth@rti.org](mailto:muth@rti.org)