International Journal of Population Data Science

Journal Website: www.ijpds.org

From Shopping Baskets to Diabetes: Causal Inference Using Retail Transaction Data

Raphael Derecki[1], Brian O'Shea[1], and James Goulding[1]

[1]N/LAB, University of Nottingham, UK

Introduction & Background

More than 3 million people in England are diagnosed with diabetes, with prevalence rising steadily and placing substantial burden on health services. Sugar consumption is an established modifiable risk factor, yet quantifying its population-level impact remains challenging; self-reported dietary intake is prone to recall bias and social desirability effects. Retail transaction data offer a novel lens into population health behaviours at scale. Aggregated shopping records provide more objective, longitudinal measures of food purchasing patterns. Supermarket loyalty card data capture millions of transactions, revealing granular variation in dietary behaviours across geographic areas. This creates opportunities to link consumption proxies to health outcomes.

Objectives & Approach

We estimate the causal effect of sugar-sweetened beverage purchasing on area-level diabetes prevalence in England. Aggregating three years of retail transaction data (2019–2021) from a major UK supermarket chain to Middle Layer Super Output Areas ($n \approx 2{,}606$), we link sugar consumption proxies to 2024 prevalence of type 2 diabetes from NHS Quality and Outcomes Framework data. We employ Double Machine Learning to flexibly adjust for area-level confounders including deprivation, demographics, food environment access, air quality, and urbanicity, with Conley spatial HAC standard errors and spatial block bootstrap at Local Authority District level confirming robustness to spatial autocorrelation and geographic aggregation. Crucially, we validate our approach using positive control outcomes, conditions with established causal pathways from sugar consumption (i.e., obesity, hypertension), and negative control outcomes, conditions with no plausible pathway (e.g., epilepsy, learning disabilities), as falsification tests for residual confounding and temporal misalignment.

Relevance to Digital Footprints

Retail transaction data are relatively well-established within the digital footprints literature for predictive modelling of health behaviours. This work extends that paradigm from prediction to causal inference, demonstrating how the same passively collected commercial data can support population-level causal claims when combined with appropriate identification strategies.

Results

Sugar purchasing shows a substantial positive effect on diabetes prevalence. Positive control outcomes display a coherent biological gradient, with effect magnitudes following the expected metabolic cascade (obesity > hypertension > diabetes). Negative control outcomes show effects an order of magnitude smaller, suggesting adequate confounding control.

Conclusions & Implications

Retail transaction data can support causal inference for diet-health relationships at population scale. The combination of positive and negative control validation strengthens causal interpretation and offers a template for future digital footprint studies where confounding is a concern.

Keywords

causal inference, retail transaction data, double machine learning, sugar-sweetened beverages, population health

https://doi.org/10.23889/ijpds.v11i4.3797