Ann. N.Y. Acad. Sci. ISSN 0077-8923

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Issue: *Paths of Convergence for Agriculture, Health, and Wealth*

A method for neighborhood-level surveillance of food purchasing

David L. Buckeridge,[1,2] Katia Charland,[1,2] Alice Labban,[3] and Yu Ma[4]

[1]Surveillance Lab, McGill Clinical and Health Informatics, [2]Department of Epidemiology, Biostatistics and Occupational Health, [3]Desautels Faculty of Management, McGill University, Montréal, Québec, Canada. [4]Department of Marketing, Business Economics, and Law, Alberta School of Business, University of Alberta, Edmonton, Alberta, Canada

Address for correspondence: David L. Buckeridge, M.D., Ph.D., McGill Clinical and Health Informatics, Montréal, Québec, H3A 1A3, Canada. david.buckeridge@mcgill.ca

Added sugar, particularly in carbonated soft drinks (CSDs), represents a considerable proportion of caloric intake in North America. Interventions to decrease the intake of added sugar have been proposed, but monitoring their effectiveness can be difficult due to the costs and limitations of dietary surveys. We developed, assessed the accuracy of, and took an initial step toward validating an indicator of neighborhood-level purchases of CSDs using automatically captured store scanner data in Montreal, Canada, between 2008 and 2010 and census data describing neighborhood socioeconomic characteristics. Our indicator predicted total monthly neighborhood sales based on historical sales and promotions and characteristics of the stores and neighborhoods. The prediction error for monthly sales in sampled stores was low (2.2%), and we demonstrated a negative association between predicted total sales and median personal income. For each $10,000 decrease in median personal income, we observed a fivefold increase in predicted monthly sales of CSDs. This indicator can be used by public health agencies to implement automated systems for neighborhood-level monitoring of an important upstream determinant of health. Future refinement of this indicator is possible to account for factors such as store catchment areas and to incorporate nutritional information about products.

Keywords: public health surveillance; nutrition; obesity; statistical methods

Introduction

The global obesity epidemic is attributable to increases in caloric intake and sedentary lifestyle,[1] but there are few effective interventions to address these fundamental causes.[2,3] A significant and growing proportion of total caloric intake is attributable to added sugar,[4] which is associated with body mass index (BMI) in both men and women.[5] In North America, sugar-sweetened drinks, such as carbonated soft drinks (CSDs), are the primary source of added sugar.[6] Interventions that decrease the intake of CSDs and other sources of added sugar can reduce caloric intake and BMI,[7] and may play a role in controlling obesity. In this context, population monitoring of added sugar intake is important so that public health agencies can assess this health determinant and evaluate the effect of interventions to decrease added sugar intake.

Added sugar intake can be measured using individual diet surveys, but these methods are resource intensive and are subject to considerable measurement error[8] and reporting bias.[9,10] Technological innovations in diet measurement offer promise for the future,[11] but individual diet measurement remains difficult to accomplish accurately in a representative and ongoing manner within constrained public health budgets. Owing to these limitations, it is not practical for public health agencies to routinely use surveys to monitor trends in diet over time at a high geographical resolution. Tracking food purchasing, however, offers a novel alternative to diet surveys for monitoring population intake of added sugar.[12,13]

Many grocery and convenience stores now use digital scanners to identify items at checkout and to generate an electronic record of sales. Companies such as Nielsen obtain these electronic sales data routinely from randomly sampled stores around the

doi: 10.1111/nyas.12332

world and make aggregated sales data and information products available to companies for marketing and other purposes.[14] These aggregate sales data should also allow routine surveillance over time and geography of purchases of sugar-sweetened products. The information derived from such surveillance may be a useful complement to the limited data available through dietary surveys and other sources.

In previous work, we developed a system of indicators from these aggregate food sales data to describe product sales by category and to account for factors such as in-store promotions.[15] In this paper, we build on our earlier work to develop a surveillance indicator that will allow tracking of neighborhood-level purchases for a food category over time. In particular, we use CSDs as an example food category due to their importance. We then validate our surveillance indicator by demonstrating its inverse association with median personal income, a relationship others have shown previously using data from dietary surveys.[16]

Methods

Data

The Island of Montreal had a population of approximately 1.6 million people in 2006. For our analyses, we partitioned Montreal into 98 forward sortation areas (FSAs). The FSA is a unit of postal geography defined by the first three digits of the six-digit postal code used in Canada. In urban areas, the size and population of an FSA is roughly equivalent to that of a ZIP code in the United States.

Data on sales of foods in 16 categories between 2008 and 2010 on the Island of Montreal were obtained from Nielsen. To select stores for sampling, Nielsen arranges all stores in each FSA into substrata based on store characteristics (e.g., total sales, square footage, and number of checkouts), and stores are then sampled randomly from each substratum. If a selected store is not scanning, a replacement store is selected from the same substratum. The selected sample is compared against all stores in the FSA, and if the sample is not representative, a partial reselection is performed. A field audit team from Nielsen visits each sampled store on a weekly basis to quantify all product display and promotion activity, and the digital sales data are obtained automatically from scanners in the same stores.

Scanner data were available as a single row for each product sold with variables to indicate the stock-keeping unit (SKU), the FSA of the store, the type of store (grocery or convenience store), the week of the sale, whether the product was on promotion through placement in the store (binary variable), the purchase price, the regular price, and whether the product was being advertised in the region (binary variable). We created the CSD category by grouping together all SKUs for flavored soft drinks containing sugar. Diet soft drinks were not included in the CSD category. Using methods we developed previously, we determined the discount frequency and advertising intensity for the CSD category.[15] Individual items were aggregated to arrive at total CSD sales by FSA and month, and corresponding summary measures for discount frequency, in-store promotion, and advertising intensity.

Stores were sampled in 76 of the 98 FSAs and data were consistent (i.e., without obvious errors such as zero stores sampled with positive sales) and regularly available for only 68 FSAs. We computed values of our indicator for these 68 FSAs with sampled stores and regularly available data. Approximately 10% of the records were missing data on sales, food category pricing, and/or marketing indicators. In order to avoid dropping incomplete observations, we imputed missing values. Data on the total number of stores by type in each FSA were obtained from the Institut national de santé publique du Québec.[17] Total population, average number of children per household, proportion speaking French or English, and median personal income were obtained for each FSA from the Canada 2006 Census. We considered the proportion of the population speaking French or English as a marker for recent immigration.

CSD indicator

An indicator of CSD purchasing should provide estimates at the level of a neighborhood so that local public health departments can target and evaluate interventions. For the sales data, however, some observations are missing over time owing to rotation of sampled stores and missing at different proportions across food categories. Consequently, the proportion of total stores and mix of stores sampled may differ by neighborhood and over time. To develop a robust surveillance indicator that is comparable over time and across geographical regions, we partitioned the data into a training set (all monthly data for each sampled neighborhood in 2008 and

Table 1. Variables in candidate forecasting models and prediction accuracy of the models for 2010 data when trained using data from 2008 and 2009

		Model			
	Source	1	2	3	4
Stores					
Number of sampled outlets	Nielsen	■	■	■	■
Number of sampled grocery stores	Nielsen	■	■	■	■
Total number of grocery stores in FSA	INSPQ	■	■	■	
Total number of convenience stores in FSA	INSPQ	■	■	■	
Sales and promotion					
Number of SKUs	Nielsen	■	■	■	■
Regular price of product	Nielsen	■	■	■	■
Discount frequency	Nielsen	■	■	■	■
In-store promotion	Nielsen	■	■	■	■
Advertising intensity	Nielsen	■	■	■	■
Sociodemographic					
Population	Census	■	■	■	
Average number of children in household	Census	■	■	■	
Proportion speaking French or English	Census	■	■	■	
Temporal and spatial					
Season indicator			■		■
Month indicator				■	
FSA random effect					■
Mean absolute prediction error		6.03%	5.75%	5.77%	2.17%

Note. The sociodemographic variable "proportion speaking French or English" is included as a marker of recent immigration.

2009) to construct a monthly sales prediction model and a test set (all monthly data in 2010 for each neighborhood) to assess the accuracy of predicted monthly sales. We first imputed missing values in the training set; then we built and assessed the accuracy of a model to predict monthly CSD sales for sampled stores in each neighborhood; then we predicted total monthly CSD sales from all stores in each neighborhood for 2010. Finally, we assessed the association at the neighborhood level between predicted total monthly CSD sales for 2010 and median personal income as measured.

Statistical analyses

The sequential regression imputation method in IVEware[18] and SAS software was used to perform multiple imputation on missing sales, pricing, and marketing data in the training sample (2008–2009) to produce five imputed data sets. In subsequent analysis, regression models were fit to each of the imputed data sets and the parameter estimates from the five models were combined using the mianalyze procedure in SAS version 8.2.[19] We compared the prediction error of four regression models, which included FSA-level variables describing stores, sales and promotions, sociodemographic characteristics, FSA indicator variables, and temporal trends (Table 1). The outcome was log-transformed monthly sales in the FSA. Overall error in predictions of monthly FSA sales for sampled stores in 2010 was measured using the mean absolute predictive error (MAPE), which is calculated as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i - \hat{Y}_i}{Y_i}, \qquad (1)$$

where n is the number of observations, Y_i is the observed data, and $\hat{Y}_i$ is the forecast.

The model with the lowest MAPE was used to predict total sales of CSDs in sampled stores for each month and FSA in 2010. To do this, we used the total number of outlets in the FSA and the total

number of grocery stores in the FSA as the inputs for the number of sampled outlets and the number of sampled grocery stores in the predictive model. Details of this model are given in the Supporting Information.

The predicted total FSA monthly log-transformed sales for 2010 were used as the outcome in assessing the relationship between median personal income and sales of CSDs. We used a Bayesian spatiotemporal model that accounts for spatially structured and unstructured variation in sales across FSAs and the serial correlation in sales through time.[20] To account for spatial autocorrelation in FSA sales, we used a conditional autoregressive prior distribution. We used additional FSA random effects for unstructured variation in sales. For the temporal correlation, we used a first-order random walk model. The FSA random effects accounted for unmeasured FSA-level confounders. Median personal income was the independent variable of interest. The spatiotemporal models were implemented using WinBUGS 1.4. We used three Markov chain Monte Carlo (MCMC) chains from different initial values to assess convergence. A detailed description of this model is available in the Supporting Information. For mapping sales data and demographic variables, observations were grouped into classes to give equal-sized classes, with each class representing a quartile.

Results

Predictive model

Using only the average of the logarithm of monthly sales of CSDs in each FSA from 2008 and 2009 to predict monthly FSA sales in 2010 gave a prediction error of 13.2%. We considered four predictive regression models in an attempt to improve on this simple model (Table 1). The best model with the lowest prediction error for sales from sampled stores (2.2%) was used to predict log-transformed total sales for each FSA and month in 2010.

Descriptive analysis

Table 2 presents descriptive statistics for the FSA with and without sampled stores. The FSA without sampled stores tends to have a lower population, but is comparable with respect to the other census variables examined. Plots of the predicted and observed sales by region over time indicated that CSD sales

Table 2. Descriptive statistics for the forward sortation areas (FSA) with and without sampled stores

	Percentiles					95% Confidence		
Variable	Min	25	Median	75	Max	Mean	Lower	Upper
FSA with sampled stores ($n = 68$)								
Monthly CSD sales ($)	454	1934	4554	20,091	117,428	15,946	10,016	21,876
Number of SKUs	24	66	76	243	349	139	115	162
Number of outlets	1.00	1.00	2.00	3.00	6.00	2.25	1.95	2.55
Number of grocery stores	0.00	0.00	1.00	1.25	4.00	0.96	0.71	1.20
Discount frequency	0.00	0.01	0.10	0.45	0.97	0.23	0.17	0.29
In-store promotion	0.00	0.00	0.00	0.22	0.97	0.13	0.09	0.18
Advertising intensity	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16
Population	5944	15,971	21,060	27,502	50,916	21,817	19,794	23,840
Median household income	14,492	20,002	23,375	26,831	44,705	24,162	22,791	25,534
Proportion speaking English or French	0.87	0.97	0.98	0.99	1.00	0.97	0.97	0.98
Average number of children in household	0.64	0.92	1.06	1.25	1.46	1.07	1.03	1.12
FSA without sampled stores ($n = 30$)								
Population	1595	5103	11,477	18,055	24,901	12,315	9742	14,889
Median household income	16,051	20,995	23,952	29,491	48,182	25,633	23,000	28,266
Proportion speaking English or French	0.88	0.98	0.98	0.99	1.00	0.98	0.97	0.99
Average number of children in household	0.48	0.76	1.02	1.22	1.44	0.98	0.88	1.08

Note. Sales and marketing data are from Nielsen, and population and household characteristics are from the Canada 2006 Census.

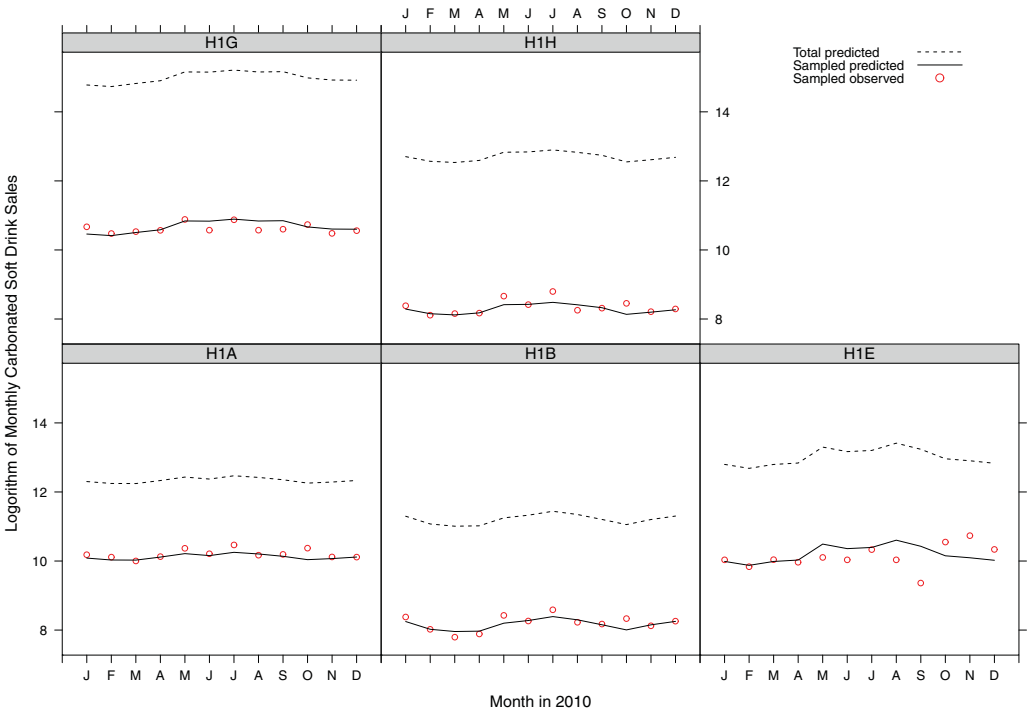

Figure 1. Predicted and observed monthly sales of carbonated soft drinks for five (H1A, H1B, H1E, H1G, and H1H) of the 68 forward sortation areas in Montreal with sampled stores. The seasonal variation is statistically significant, as indicated by the parameter estimate for the season variable (summer versus winter) of 0.212 (95% CI 0.080–0.345) in our space–time prediction model.

tend to increase in the summer within each region but that sales are relatively constant within a region over time, and considerable differences in total sales are seen across regions (Fig. 1). The spatial distribution of total predicted monthly sales indicates some spatial clustering of regions with high sales, and similar clustering is seen in median personal income (Fig. 2).

Spatial regression

Convergence of the Bayesian model was achieved following 20,000 iterations. An additional 20,000 iterations were used to estimate the random effects and regression coefficients. The regression coefficient for FSA median personal income was −0.0001641 (credible interval −0.00023, −0.000059). The logarithm of total sales ranged from 7.9 to 16.2, whereas FSA median personal income ranged from $14,273 to $37,903. The estimated coefficient for median income implies that a $10,000 decrease in FSA median income is associated with an average increase of 1.641 units in log (total sales) or an increase in total sales by a factor of exp (1.641) = 5.16.

Discussion

Using automatically captured data on monthly sales of food products, we developed an indicator of CSD sales by neighborhood. We then took an initial step toward validating this indicator of added sugar intake by demonstrating its negative correlation with median household income, a relationship that others have observed using data from a dietary survey.[16] More specifically, we found at the neighborhood level that a decrease of $10,000 in median personal income was associated with a fivefold increase in sales of CSDs. Given the importance of CSDs as a source of added sugar intake[6] and the need to identify effective public health interventions for reducing the intake of added sugar,[7,21] the ability to routinely monitor this upstream determinant of obesity can provide guidance for targeting and evaluating public health interventions. The indicator

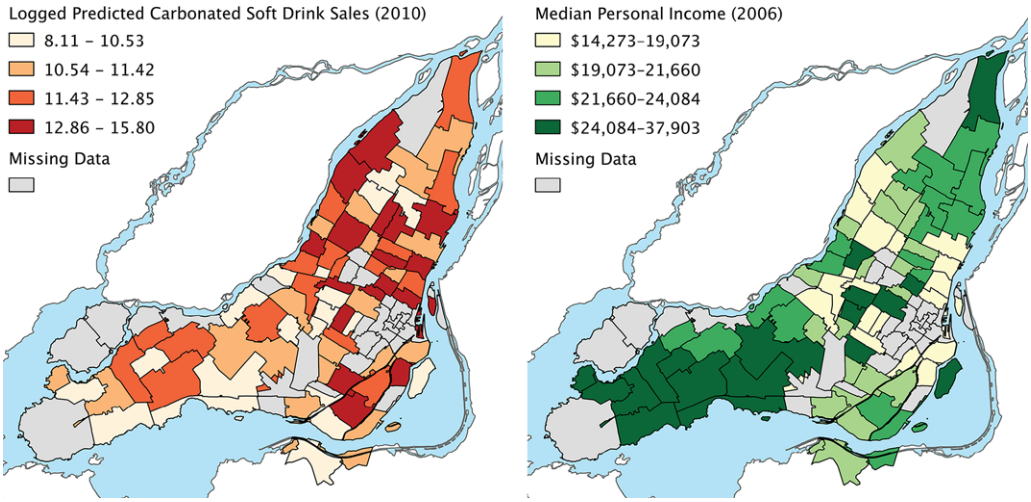

Figure 2. Geographical distributions of total predicted carbonated soft drink sales and median personal income by forward sortation areas in Montreal.

that we developed is based on data captured automatically for inventory and marketing purposes. These data are available throughout the world,[14] so it should be possible to implement this indicator in other settings with few modifications. Moreover, the automated nature of the data capture should enable the development of an automated surveillance system based on this and other similar indicators of different food categories of public health importance.[22]

The development of our surveillance indicator builds on our previous work to develop indicators from automatically captured store sales data.[12,15,22] In this work, we have demonstrated how these sales indictors can be used to develop a public health surveillance indicator, which provides interpretable information over time and at a high geographical resolution. We do not know of any previous efforts to develop a similar sort of indicator using food sales data, but in public health settings, surveillance systems increasingly rely on automated feeds of data captured for other purposes, such as electronic medical records,[23] telehealth calls,[24] and pharmaceutical sales.[25] With respect to our substantive finding, the small amount of evidence regarding the negative association between CSD sales and median personal income is based on survey data and is difficult to compare to our results, as the measures of added sugar intake and income were different in those studies.[16,26] For example, one study reported a 15% increase in added sugar intake when comparing the third of respondents with the highest family income to the third of respondents with the lowest family income.[16]

Although we have identified a novel approach to monitoring food purchasing at the neighborhood level, there are limitations to this approach. For one, the indicator measures food purchasing, not dietary intake. However, we did demonstrate that our indicator of CSD sales has an inverse association with income, as reported by others. Nonetheless, because our indicator follows purchasing at the neighborhood level, it is not possible to attribute purchases to specific population subgroups, although differences in demographic profiles between neighborhoods could be used to assess ecological associations between population subgroups and purchasing patterns. Another limitation is that our indicator measures a single food category and not a variable of more direct nutritional interest, such as added sugar. In the case of added sugar, CSDs represent a large proportion of the total intake,[6] so the connection is clear. In general, however, we could address this limitation in the future by linking data on nutritional values of products to data on sales[13] and developing new indicators of total added sugar or other measures, which would be the product of the sales data and the nutritional value of each item sold. Monitoring multiple categories of food purchasing simultaneously may also be important to identify

if interventions are resulting in substitution of one food category for another.[12] The trend toward dining out creates another limitation to our approach. The indicator measures only food purchased by consumers for eating at home and does not capture sales of food at restaurants or purchasing from restaurants. In the United States, the frequency of dining out increases with family income,[27] but the energy density of restaurant meals is inversely associated with family income.[28] There is also some spatial imprecision inherent in the estimates, as we directly attribute the sales information from stores to the regions that they fall within, making no allowance for store catchment areas. In future work, it should be possible to address this limitation by defining store catchment areas and then using those areas to assign sales proportionally to regions. Finally, we relied on multiple imputation to estimate missing values in the sales data. Only 10% of the values were missing in our data, but if a large proportion of values were missing, then caution would be warranted in using this approach.

The method that we have developed for monitoring food purchasing opens many avenues for future research. One avenue of research is to refine and extend the method. Particularly useful extensions would be to incorporate catchment areas for stores and to link food products to nutritional databases, allowing population-level monitoring of purchasing at the nutrient level. Another avenue of research is to extend the method to multiple food categories, allowing near real-time monitoring of the full breadth of food purchased within neighborhoods. This comprehensive view would allow a richer understanding of how the complete food "basket" varies over time and across neighborhoods. Perhaps the most promising avenue of research is using this method, ideally extended to measure nutrition or the full basket, to discover the effect of interventions and neighborhood-level characteristics on spatial and temporal variations in food purchasing.

In conclusion, we have developed, assessed the accuracy of, and begun to validate an indicator to allow neighborhood-level surveillance of CSD purchasing. This indicator should be straightforward to implement in other settings, and there are many ways that this indicator can be refined and extended in the future to support automated surveillance of food purchasing within neighborhoods.

Conflicts of interest

The authors declare no conflicts of interest.

Supporting Information

Additional supporting information may be found in the online version of this article.

Detailed description of regression models

References

1. World Health Organization. 2000. Obesity: Preventing and managing the global epidemic. Report of a WHO consultation. *World Health Organ. Tech. Rep. Ser.* **894:** i–xii, 1–253.
2. Epstein, L.H. *et al.* 2012. Experimental research on the relation between food price changes and food-purchasing patterns: a targeted review. *Am. J. Clin. Nutr.* **95:** 789–809.
3. Metcalf, B., W. Henley & T. Wilkin. 2012. Effectiveness of intervention on physical activity of children: systematic review and meta-analysis of controlled trials with objectively measured outcomes (EarlyBird 54). *Br. Med. J.* **345:** e5888.
4. Welsh, J.A. *et al.* 2010. Caloric sweetener consumption and dyslipidemia among US adults. *J. Am. Med. Assoc.* **303:** 1490–1497.
5. Wang, H., L.M. Steffen, X. Zhou, *et al.* 2013. Consistency between increasing trends in added-sugar intake and body mass index among adults: the Minnesota heart survey, 1980–1982 to 2007–2009. *Am. J. Public Health* **103:** 501–507.
6. Report of the Dietary Guidelines Advisory Committee on the Dietary Guidelines for Americans. 2010. <http://www.cnpp.usda.gov/dgas2010-dgacreport.htm>.
7. Ebbeling, C.B. *et al.* 2012. A randomized trial of sugar-sweetened beverages and adolescent body weight. *N. Engl. J. Med.* **367:** 1407–1416.
8. Natarajan, L. *et al.* 2010. Measurement error of dietary self-report in intervention trials. *Am. J. Epidemiol.* **172:** 819–827.
9. Salvini, S. *et al.* 1989. Food-based validation of a dietary questionnaire: the effects of week-to-week variation in food consumption. *Int. J. Epidemiol.* **18:** 858–867.
10. Sallé, A., M. Ryan & P. Ritz. 2006. Underreporting of food intake in obese diabetic and nondiabetic patients. *Diabetes Care* **29:** 2726–2727.
11. Illner, A.-K. *et al.* 2012. Review and evaluation of innovative technologies for measuring diet in nutritional epidemiology. *Int. J. Epidemiol.* **41:** 1187–1203.
12. Ma, Y., K.L. Ailawadi & D. Grewal. 2013. Soda versus cereal and sugar versus fat: drivers of healthful food intake and the impact of diabetes diagnosis. *J. Mark.* **77:** 101–120.
13. Ng, S.W. & B.M. Popkin. 2012. Monitoring foods and nutrients sold and consumed in the United States: dynamics and challenges. *J. Acad. Nutr. Diet.* **112:** 41–45.e4.
14. Nielsen Corporation. http://www.nielsen.com/us/en.html. Accessed October 16, 2013.
15. Ma, Y. *et al.* 2013. System of Indicators for the Nutritional Quality of Marketing and Food Environment: Product Quality, Availability, Affordability, and Promotion. In *Diet Quality: An Evidence Based Approach*, Volume 2. V.R. Preedy, Ed.: Springer. New York.

16. Thompson, F.E. *et al.* 2009. Interrelationships of added sugars intake, socioeconomic status, and race/ethnicity in adults in the United States: National Health Interview Survey, 2005. *J. Am. Diet. Assoc.* **109:** 1376–1383.
17. Portrait de l'environment bâti et de l'environment des services. (Institut national de santé publique du Québec). <http://environnementbati.inspq.qc.ca>.
18. Raghunathan, T.E., J.M. Lepkowski, J. Van Hoewyk & P. Solenberger. 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Stat. Can.* **27:** 85–95.
19. The MIANALYZE Procedure. <http://support.sas.com/rnd/app/papers/mianalyzev802.pdf>.
20. Knorr-Held, L. & J. Besag. 1998. Modelling risk from a disease in time and space. *Stat. Med.* **17:** 2045–2060.
21. Kaiser, K.A., J.M. Shikany, K.D. Keating & D.B. Allison. 2013. Will reducing sugar-sweetened beverage consumption reduce obesity? Evidence supporting conjecture is strong, but evidence when testing effect is weak. *Obes. Rev.* **14:** 620–633.
22. Buckeridge, D.L. *et al.* 2012. An infrastructure for real-time population health assessment and monitoring. *IBM J. Res. Dev.* **56:** 2:1–2:11.
23. Greene, S.K. *et al.* 2012. Gastrointestinal disease outbreak detection using multiple data streams from electronic medical records. *Foodborne Pathog. Dis.* **9:** 431–441.
24. Harcourt, S.E. *et al.* 2001. Can calls to NHS Direct be used for syndromic surveillance? *Commun. Dis. Public Health* **4:** 178–182.
25. Bounoure, F., P. Beaudeau, D. Mouly, *et al.* 2011. Syndromic surveillance of acute gastroenteritis based on drug consumption. *Epidemiol. Infect.* **139:** 1388–1395.
26. Haley, S., J. Reed, B.H. Lin & A. Cook. 2005. *Sweetener Consumption in the United States: Distribution by Demographic and Product Characteristics.* Economic Research Service, US Department of Agriculture.
27. Kwon, Y., S. Oh, S. Park & Y. Park. 2010. Association between household income and overweight of Korean and American children: trends and differences. *Nutr. Res.* **30**: 470–476.
28. Bezerra, I.N., C. Curioni & R. Sichieri. 2012. Association between eating out of home and body weight. *Nutr. Rev.* **70:** 65–79.