



Demographic, spatial and temporal dietary intake patterns among 526 774 23andMe research participants

Janie F Shelton*, Briana Cameron, Stella Aslibekyan, 23andMe Research Team and Robert Gentleman

23andMe, Inc., Sunnyvale, CA 94086, USA

Submitted 14 May 2019; Final revision received 19 March 2020; Accepted 6 April 2020; First published online 29 June 2020

Abstract

Objective: To characterise dietary habits, their temporal and spatial patterns and associations with BMI in the 23andMe study population.

Design: We present a large-scale cross-sectional analysis of self-reported dietary intake data derived from the web-based National Health and Nutrition Examination Survey 2009–2010 dietary screener. Survey-weighted estimates for each food item were characterised by age, sex, race/ethnicity, education and BMI. Temporal patterns were plotted over a 2-year time period, and average consumption for select food items was mapped by state. Finally, dietary intake variables were tested for association with BMI.

Setting: US-based adults 20–85 years of age participating in the 23andMe research programme.

Participants: Participants were 23andMe customers who consented to participate in research (n 526 774) and completed web-based surveys on demographic and dietary habits.

Results: Survey-weighted estimates show very few participants met federal recommendations for fruit: 2.6 %, vegetables: 5.9 % and dairy intake: 2.8 %. Between 2017 and 2019, fruit, vegetables and milk intake frequency declined, while total dairy remained stable and added sugars increased. Seasonal patterns in reporting were most pronounced for ice cream, chocolate, fruits and vegetables. Dietary habits varied across the USA, with higher intake of sugar and energy dense foods characterising areas with higher average BMI. In multivariate-adjusted models, BMI was directly associated with the intake of processed meat, red meat, dairy and inversely associated with consumption of fruit, vegetables and whole grains.

Conclusions: 23andMe research participants have created an opportunity for rapid, large-scale, real-time nutritional data collection, informing demographic, seasonal and spatial patterns with broad geographical coverage across the USA.

Keywords
Nutrition surveys
Dietary factors
23andMe
BMI

Inadequate intake of healthy foods (fruits, vegetables, whole grains and nuts/seeds) and excessive intake of foods high in Na, added sugars and saturated fat are major contributors to excess morbidity and premature mortality worldwide^(1,2). In the USA, approximately half of all cardio-metabolic deaths are attributable to suboptimal intake of fruits, vegetables, nuts, seeds and whole grains and excess consumption of salt, processed meats and sugary beverages⁽³⁾ and per the dietary guidelines outlined by the American Heart Association to maximise cardiovascular health, 45.6 % of Americans adults are estimated to have a 'poor diet', with <10 % consuming adequate amounts of fruits and vegetables⁽⁴⁾.

Poor dietary habits and inadequate physical activity are major drivers of elevated BMI, which in turn increases the risk of developing adverse cardio-metabolic outcomes⁽⁵⁾. Diet and physical activity also represent the most actionable areas at both the individual and population levels to prevent chronic disease⁽⁶⁾. While the majority of dietary recommendations are geared towards average effects with measurable benefits to population health, the science of precision nutrition has been uncovering informative subgroup variation.

For example, among those who consume 22 g/d or more of saturated fat, weight gain was more pronounced among those with the -265 C/C genotype of *APOA2*, an

*Corresponding author: Email jshelton@23andme.com



estimated 10–20% of the population, compared with those without it⁽⁷⁾. However, longer term health consequences, such as varying cardio-metabolic disease risk across combinations of exposures, are not well understood. While precision nutrition has been described as a major priority for epidemiology⁽⁸⁾, the vast majority of studies published to date are underpowered for more granular discovery.

In order to advance our understanding of how nutritional and other factors interact and impact health, data are needed from large populations followed over time. Since 2017, over 500 000 genotyped 23andMe research participants answered a survey about their dietary habits over the past month. Due to the growing size of the customer database, widespread geographical representation and continuous data collection over time, data provided by 23andMe research participants now represent a large enough sample to inform population-based inferences for a variety of health behaviours, including diet.

However, because 23andMe participants are a subsample of 23andMe personal genome service customers, they are not a representative sample of the general population. Therefore, the aim of this manuscript is to describe how the sociodemographic profile and dietary habits of the 23andMe research participants compare with the US population, to characterise its dietary habits using survey weights that account for potential imbalances and to test cross-sectional associations between BMI and consumption of several food items in this uniquely large cohort. Our manuscript illustrates both the potential and the caveats of conducting nutritional epidemiology research in large-scale digital cohorts.

Methods

This study used data from consented 23andMe research participants from the US aged 20 years or older who completed the National Health and Nutrition Examination Survey (NHANES) 2009–2010 dietary screener questionnaire on the 23andMe website or mobile application.

This study was conducted according to the guidelines laid down in the Declaration of Helsinki, and all procedures involving research study participants were approved by the Ethical and Independent Review Services, a private independent institutional review board. All 23andMe customers consent to participate in research online, and consent is captured electronically. Ethical and Independent Review Services approved this form of informed consent and waived the requirement to obtain signed consent under US law 45 CFR 46.117(c).

Customer data are de-identified; 23andMe researchers who conducted the statistical analyses in this manuscript did not have access to personally identifiable information (e.g. name and address) and were trained in the responsible conduct of research. 23andMe has obtained a Certificate of Confidentiality from the National Institutes

of Health, further protecting the privacy of research participants. Additional consent form information is available at www.23andme.com/about/consent.

Data collection

Once a 23andMe personal genome service customer receives their sample collection kit, they are asked to register it online prior to return. During this process, all customers are invited to participate in research, which occurs predominantly through web-based research surveys, some of which are developed in-house by 23andMe and others are implemented using previously validated instruments. Beginning in March 2017, dietary intake frequencies were collected using a validated web-based version of the self-administered 25-item NHANES dietary screener questionnaire (DSQ) (2009/2010)⁽⁹⁾. Accompanying surveys provided self-reported data on covariates such as age, sex, race/ethnicity, education and BMI. The questions on race/ethnicity and education utilised the same response options as the US Census. Participant recruitment to the surveys took place through two approaches, actively through email and passively through the website and mobile application. Passive recruitment on the website occurs via the 'research stream', which is a feature on the 23andMe website and within the 23andMe app (available on both iOS and Android devices) which continuously surfaces surveys to eligible respondents on a variety of topics.

As such, the DSQ was fielded to those who had likely already completed higher priority surveys such as the Health survey (which collects data on basic demographics, health and disease status), or disease-specific surveys based on conditions they may have reported. Because of this passive recruitment targeting scheme, we do not know the total population of people who were offered the survey, and therefore cannot directly estimate a survey response rate.

Dietary assessment and intake estimation

The DSQ measures intake of fruits and vegetables (cups/d), dairy (cups/d), Ca (mg), added sugars (g), whole grains (ounce equivalents) and fibre (g)⁽¹⁰⁾. Total fruit includes both whole fruit and fruit juice, vegetable intake is estimated from consumption frequency of salad, potatoes, beans, tomato sauce, salsa, pizza toppings and other vegetables. Dairy intake is estimated from milk, cheese, ice cream and pizza. Whole grains are derived from cereals, whole-grain bread, popcorn and whole-grain rice. Added sugars are derived from soda, fruit drinks, cookies/cakes/pie, doughnuts, ice cream, added sugar/honey in coffee or tea, candy and cereal. Added sugars from sugar-sweetened beverages are derived from soda, fruit drinks and sugar/honey added to coffee or tea⁽¹⁰⁾.

The DSQ and scoring algorithms to derive broader food groups were validated against dietary intake using the 24-h recall method in a representative population of



non-institutionalised US-based NHANES study participants aged 2–69 years⁽¹¹⁾. In that context, the DSQ was reported to produce stable estimates of intake for this set of dietary factors and correlate well with 24-h recall estimates in three external study populations⁽¹¹⁾. In the DSQ, individual food items are ascertained as frequency over the past month. Based on the data processing and scoring procedures recommended for use with the DSQ, we converted monthly estimates to daily estimates and then multiplied the latter by item-specific portion and serving size estimates provided according to age and sex⁽¹²⁾. Components of cereals such as sugar, fibre and whole-grain content were derived from the classifications on a per cereal basis provided by NHANES.

To evaluate the percent of the population meeting dietary recommendations, we estimated the proportion of participants within strata of age and sex who met the United States Department of Agriculture's Dietary Guidelines for Americans 2015–2020 recommended intake of fruits and vegetables⁽¹³⁾.

Time trends and maps

Mean daily intake and 95 % CI were plotted by week of survey completion to explore seasonal variation in dietary intake. Although the data collected refer to the past month, we did not apply a lag period or adjust the date in any way. Broader temporal changes in reported dietary habits over time are visualised with loess curves.

Participants self-reported their current zip code, which we subsequently mapped to states, to characterise the geographic distribution of average intake of food groups, select food items and average BMI across the US. Regional average intake was estimated for the South (Delaware, District of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia, Alabama, Kentucky,

Mississippi, Tennessee, Arkansas, Louisiana, Oklahoma and Texas), West (Arizona, Colorado, Idaho, New Mexico, Montana, Utah, Nevada, Wyoming, Alaska, California, Hawaii, Washington and Oregon), Northeast (Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, New Jersey, New York and Pennsylvania) and the Midwest (Indiana, Illinois, Michigan, Ohio, Wisconsin, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota and South Dakota).

For this analysis, the foods selected for seasonality were those with the most pronounced seasonal trends: chocolate, fruit, ice cream and salad. Those selected for mapping showed the most pronounced geographic patterns: fruit, vegetables, whole grains, red meat, processed meat, dairy and pizza.

Development of survey weights

To address differences between our respondent population and the national population (Table 1), we developed survey weights and applied them to our sample. We used iterative proportional fitting to calculate weights based on age (20–64, 65+), sex (M, F), race (white, non-white), education (less than college, completed college) and BMI (obese, not obese). We weighted our sample proportion to match the demographic distribution as reported by the US Census (age, sex, race, education)⁽¹⁴⁾ and the Centers for Disease Control and Prevention⁽¹⁵⁾. We excluded respondents (n 156, 053) from the raking procedure (implemented via the survey package (version 3.29.5)⁽¹⁶⁾ in R) and subsequent analyses if they did not provide this demographic information. To address the increase in SE that may be introduced with large sample weights, we chose binary classifications instead of utilising all strata (e.g. for BMI) and trimmed all initial weights to be <5 times the mean survey weight⁽¹⁷⁾.

Table 1 Respondent characteristics with complete data on age, sex, education, race/ethnicity and BMI compared with the national population drawn from the US Census

		Unweighted sample n	%	Weighted sample n	%	US Population n (%)
Sex	Female	342 325	64.9	264 441	50.2	50.2
	Male	184 449	35.1	262 334	49.8	49.8
Age	20–64	392 580	74.5	444 597	84.4	84.4
	65–85	134 194	25.5	82 176	15.6	15.6
Race	White	453 705	86.1	403 509	76.6	76.6
	Black	7142	1.4	12 248	2.3	13.4
	American Indian	1893	0.4	4098	0.8	1.3
	Asian	11 424	2.2	15 142	2.9	5.8
	Oceanian	418	0.007	854	0.16	0.2
	Multiple	52 192	9.9	90 923	17.3	0.3
	Non-Hispanic	493 444	93.7	484 643	92.0	81.9
Ethnicity	Hispanic	33 330	6.3	42 130	8.0	18.1
	Less than college	217 521	41.3	367 320	69.7	69.7
Education	College or more	309 253	58.7	159 453	30.3	30.3
BMI	Underweight	6872	1.3	6435	1.2	1.6
	Normal weight	179 066	34.0	151 831	28.8	28.9
	Overweight	178 894	34.0	176 761	33.6	33.1
	Obese	161 942	30.7	191 746	36.4	36.4



Dietary intake associations with BMI

To measure the cross-sectional correlation between dietary intake and BMI, we explored the association between each food item continuously using the daily frequency or estimated quantity for derived measures, but ultimately classified intake by tertiles for ease of interpretation and standardisation of quantity (high- *v.* low-intake frequency). We limited our sample to participants who self-reported height and weight values yielding estimated BMI scores between 14 and 70. We evaluated BMI both with and without log transformation.

During model development, race/ethnicity, education, age, sex, age \times sex and age squared were evaluated as potential confounders. The final linear models tested the associations between food intake frequency (highest *v.* lowest consumption tertiles for all food items measured in the DSQ) and log-transformed BMI, adjusting for race/ethnicity, education, age (centred at 50 years), sex and centred age squared to maximise the variance explained by the model and the uniformity of the plotted residuals. We estimated regression coefficient estimates and 95 % CI for each food item to evaluate the relationship between frequency of intake and BMI.

We used R (R Foundation for Statistical Computing, version 3.2.5) for all statistical analyses and data visualisation.

Results

Respondent characteristics

A total of 526 774 US-based respondents with non-missing data for sex, age, education, race/ethnicity and BMI completed the DSQ between March 2017 and August 2019. Compared with the US population, sample respondents were more likely to be female (65 % *v.* 50 %), more likely to be white (86 % *v.* 77 %), less likely to be Hispanic (6 % *v.* 18 %), nearly twice as likely to have completed college (59 % *v.* 30 %) and less likely to be obese (31 % *v.* 36 %). Application of sampling weights yielded a more representative sample based on age, sex, education and BMI.

Temporal characteristics of dietary intake

Due to the high rate of gift giving of the 23andMe genetic testing kit during the holidays, survey completion between late December and early January is on average 4–5 fold greater than the average of other weeks throughout the year. Due to this high degree of response in these months, the precision of the estimates by season is highest over the Northern Hemisphere winter period as compared with other seasons (Fig. 1).

Because survey data have been collected continuously over 2 years, seasonal trends in dietary frequencies are observable. Fig. 2 shows the unadjusted mean daily intake

patterns for fruit, vegetables, salad, chocolate and ice cream. Chocolate and ice cream show clear seasonal patterns, with peak chocolate consumption (0.42 times/d) in December–January compared with the lowest consumption in mid-June of 0.31 times/d. Peak ice cream consumption is observed in June–August, with an average frequency of 0.17 times/d compared with 0.01 times/d in February.

Milk consumption, which includes both soyabean and dairy, declined by approximately 28 % between 2017 and 2019 following a pattern evident since the 1970s⁽¹⁷⁾ and observed in the NHANES study between 1992 and 2001⁽¹⁹⁾. However, only a very minimal decline in total dairy consumption is observed over time, likely owing to observed increases in cheese and pizza consumption (see online supplementary material, Supplemental Fig. S1).

For fruits and vegetables, we observed higher reported consumption in Northern Hemisphere summer than winter (Fig. 2), but also noted a general decline over the reporting period (Fig. 3). We have explored various possible explanations for this observation, such as changing customer demographics (e.g. age, sex or type of genotyping kit purchased or the influx of winter customers), but saw no clear explanatory patterns. Further, because national published estimates are not yet available for this time period, we have no comparison on which to support or refute the observation that fruit and vegetable consumption is declining in the general population.

Spatial characteristics of dietary intake

The respondent population show high geographical coverage across the USA (Fig. 1), with a minimum of 1000 respondents in every state. Higher proportions of respondents were from high population density states (California, Texas and Florida). By exploring the fraction of respondents as a percent of the total adult population, we observed relative overrepresentation in California, Florida and Pennsylvania and underrepresentation in Nebraska, South Dakota and Vermont.

Minor regional differences are noted when comparing data in aggregate (Table 4), but intake frequencies plotted at the state level show distinct dietary patterns by region. In the Southeastern states, fruit, vegetable and whole-grain intake is markedly lower than in coastal areas, and intake of processed meat, regular soda and added sugars is comparatively higher. Higher average intake of milk, pizza and red meat is observed in the northern Midwestern states, and while lower vegetable intake is observed, fruit and whole-grain intake is similar to coastal areas (Fig. 3). Dietary frequencies that correspond to higher BMI in Fig. 4 are also demonstrated to spatially correspond with patterns of higher BMI at the state level (e.g. higher frequency of processed meat and lower consumption of fruits, vegetables and whole grains in the southern US where BMI is the highest).

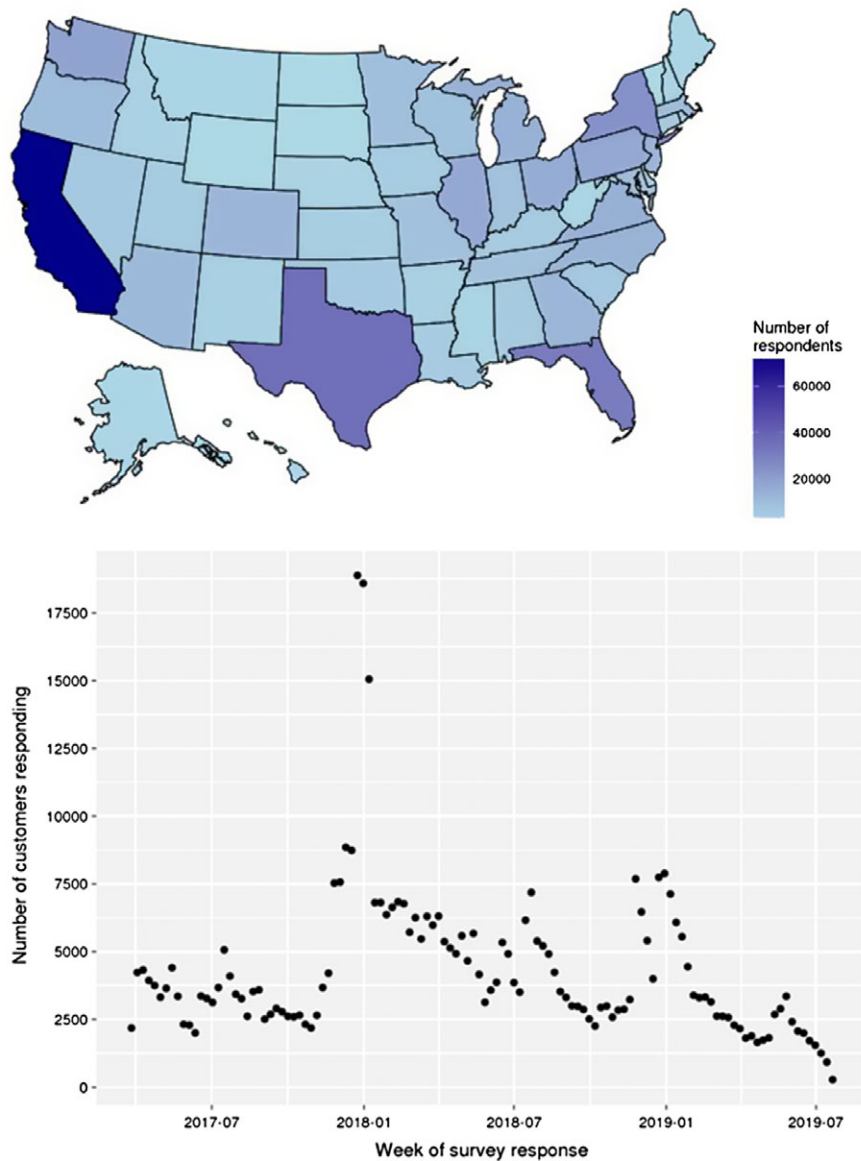


Fig. 1 Survey completions by state (A) and week (B) during interval of data collection

Dietary intake estimates

Only 2.6 % of the respondent population met national recommendations of fruit intake (1-1/2–2 cups/d for adults depending on age and sex), with an average population intake of 0.83 cups/d⁽²⁰⁾. Mean vegetable intake (excluding French fries but including other potatoes) across all respondents was 1.54 cups/d, translating to only 5.9 % of the total adult population who met guidance of 2-1/2–3 cups/d depending on age and sex (Tables 2 and 3). Vegetable intake showed a clear dose–response relationship with age, whereby among males and females aged 51 years and over, 6.1 % and 10.8 % met recommendations, respectively. Among 20–30-year-olds, only 1.1 % of males and 0.3 % of females met recommendations (Table 3)⁽²¹⁾.

In stratified estimates of average intake, college education *v.* no college education conferred higher intake of

fruits, vegetables and whole grains, with substantially less sugar intake both in total added sugars and the percent consuming at least one sugar-sweetened beverage/d. Differences across race/ethnicity showed Oceanian populations (i.e. Americans with origins in the Pacific islands) eat substantially less whole grains compared with other groups, and nearly half of American Indian respondents reported at least one sugar-sweetened beverage/d. Dairy consumption was lower among Asians compared with other groups.

As shown in Fig. 5, consumption of many of the food items included in the NHANES dietary screener shows strong associations with BMI. In models adjusted for age, sex, education and race/ethnicity, foods associated with the greatest increase in BMI included processed and red meat, fried potatoes and pastry. Higher consumption of

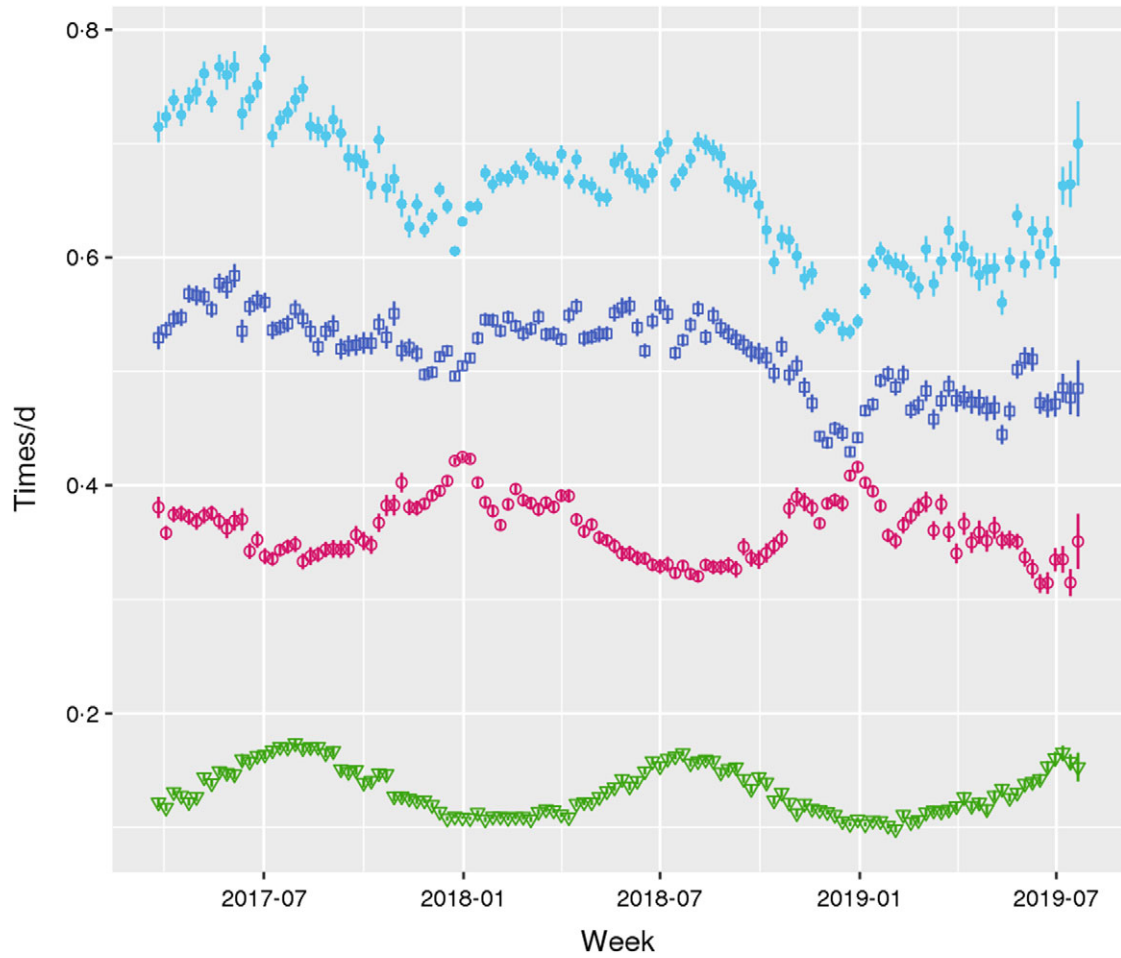


Fig. 2 Seasonality patterns of chocolate, fruit, ice cream and salad average daily intake by week of survey completion, 2017–2019. Food item: ●, chocolate; ▲, fruit; ▲, ice cream; ■, salad

dairy products (cheese and ice cream) was all associated with higher BMI. Conversely, each tertile increase in the past month consumption of fruits, vegetables, whole grains, beans and cereals corresponded to lower BMI. Frequency of chocolate consumption showed no statistically significant association with BMI.

Discussion

Here, we detail demographic, temporal and spatial characteristics of dietary factors measured by the DSQ and how they relate to BMI among 23andMe research participants. Because 23andMe research participants are twice as likely to have a college education, more likely to be female, white and older than the general US population, dietary habits differential across these characteristics were most changed by weighting the sample to better represent the broader free living US population of adults. In the weighted sample, the majority of 23andMe participants fall well below dietary guidelines for consumption of fruits and vegetables showing remarkable similarity to the broader US population. In

our stratified analyses, the most pronounced difference by education, which we used as a proxy for socio-economic status, was for soda consumption, which is less common as educational attainment increases. In our cross-sectional analyses of BMI and dietary habits, these under-consumed foods corresponded to lower BMI, whereas dairy, meats and added sugars were associated with higher BMI. Descriptive spatial patterns of models of high *v.* low intake of select dietary factors and BMI replicate results from large-scale nationally representative studies such as the Behavioral Risk Factor Surveillance System^(23,24). Although it is generally agreed that there exists a troubling degree of bias in self-reported dietary intake data⁽²⁵⁾, observation of these patterns which are replicated across more rigorous data collection methods⁽²⁶⁾ are reassuring.

Direct comparisons between weighted intake estimates among 23andMe research participants and the NHANES sample may be limited based on methodological differences in data collection. As described, 23andMe estimates are based on a brief screener, whereas the NHANES estimates are based on two in-person 24-h food intake interviews⁽²²⁾. Comparing the survey-weighted results to

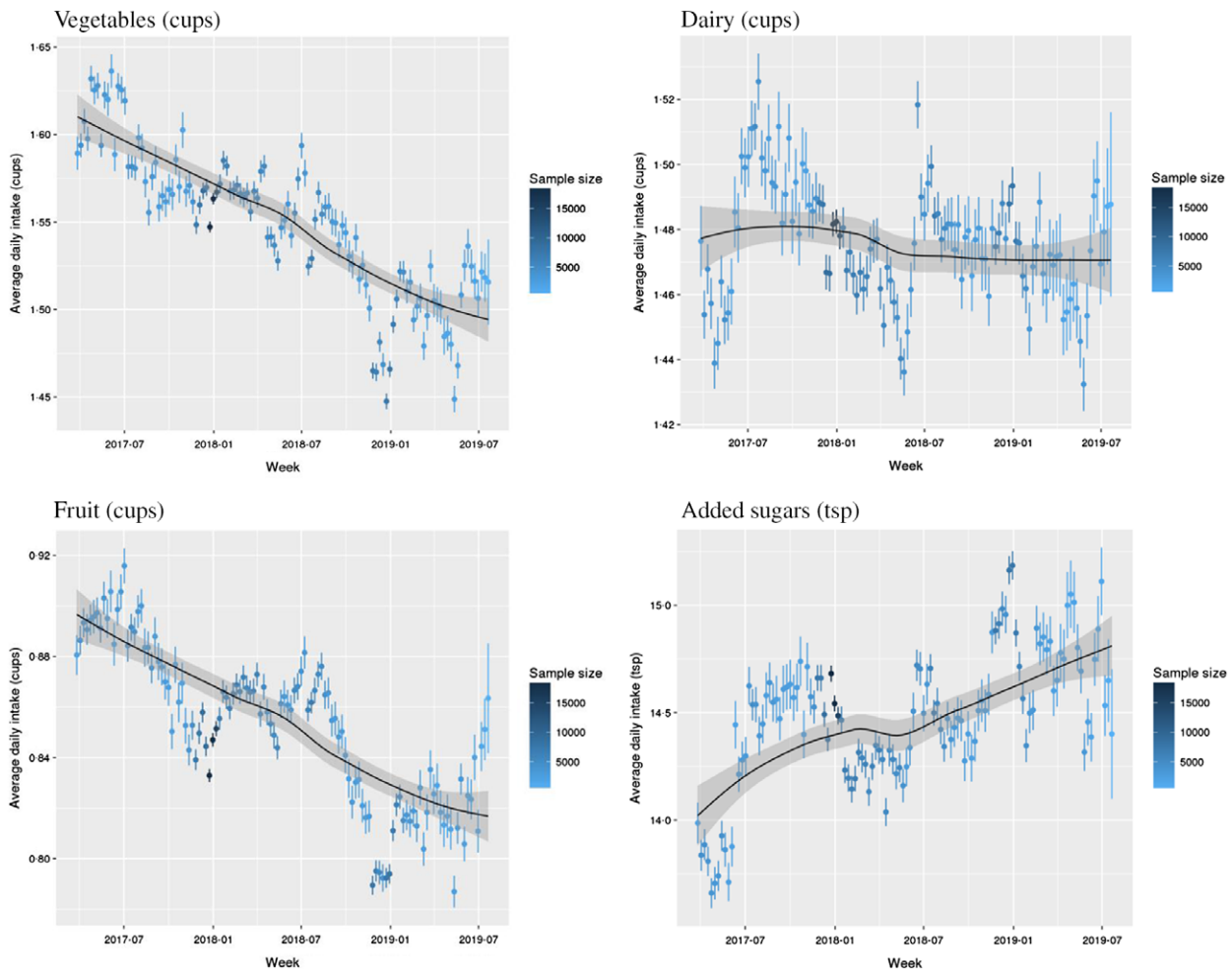


Fig. 3 Mean and 95 % CI with loess curves of the of intake frequency reported by week of data collection for vegetables, fruit, dairy and added sugars, over the 2-year data collection period (2017–2019)

the 2011–2012 NHANES study, 23andMe respondents were most similar to the general US population for fruits (0.73 *v.* 0.69 median cups/d) and (1.54 *v.* 1.63 cups/d) vegetables but less so for added sugars (15.4 *v.* 18.5 tsp/d) and whole grains (0.56 *v.* 1.00 ounces/d). Because of the significant time gap (5–6 years) between the most recently published NHANES values and the 23andMe data, it is possible that some amount of this difference is explained by the previously reported downward trends in consumption of added sugars observed over time⁽⁴⁾.

Measurement of dietary habits in the context of a self-reported web-based survey has the benefit of efficiency, but is hindered by the limited scope of data collection and its accuracy in reflecting the individuals' general habits. Because web-based survey completion tends to decline as survey length increases⁽²⁷⁾, we chose the DSQ rather than a longer assessment such as a full food frequency questionnaire to maximise utility of the data collected with minimal respondent drop-off. However, the NHANES dietary screener (2009/2010)^(9,10) does not allow for estimation of total energy intake and may be a suboptimal tool for

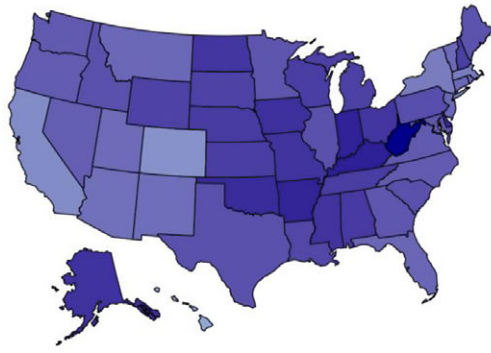
measuring the proportion of population meeting national fruit intake guidelines.

For example, due to a truncated response option at two or more times/d, for many adults (all men and women aged 19–30 years), it is not possible to report eating enough whole fruit to meet current dietary recommendations (2 cups/d for women aged 19–30 years and men aged 19–50 years) if the portion size estimates are applied, which are <1 cup per serving. This becomes a greater problem for women, because the quantity multiplier (i.e. the estimated number of cups per serving) used to estimate cups/d is smaller for women than it is for men. Because the total fruit estimate is based on the combination of whole fruit and 100 % fruit juice, participants that report the maximum serving size frequency (two or more times/d) must also drink 100 % fruit juice regularly to meet the recommended minimum fruit intake requirements.

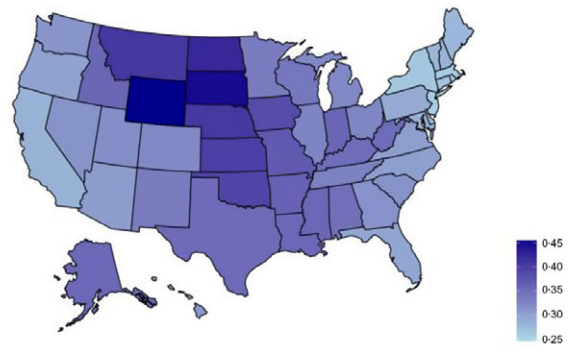
Because fruit juice is not universally considered a healthful dietary option due to its high added sugar and low fibre content⁽²⁸⁾, those who eat fruit in quantities equal to or exceeding 2 cups/d but abstain from drinking fruit



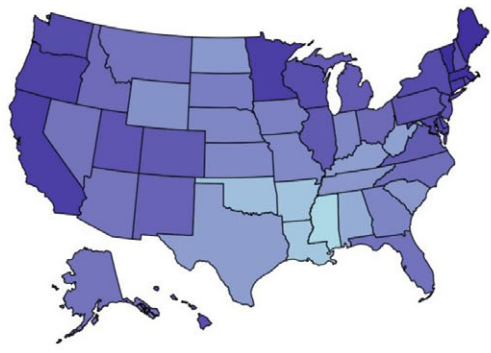
Mean BMI



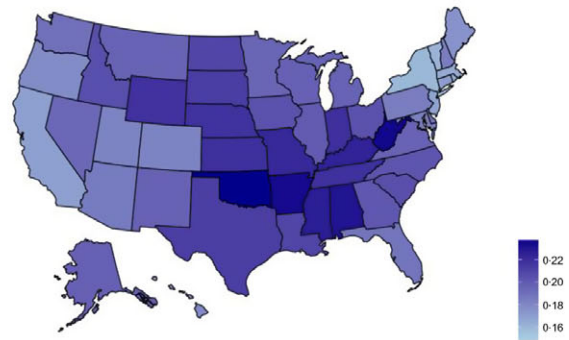
Mean daily intake of red meat



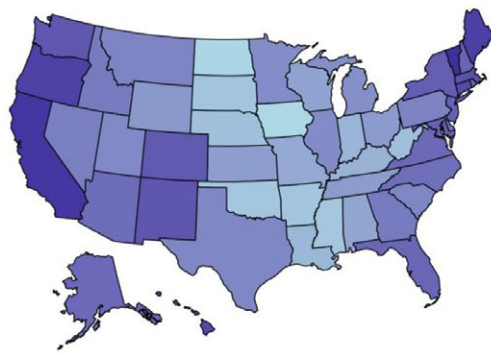
Mean fruit intake (cups)



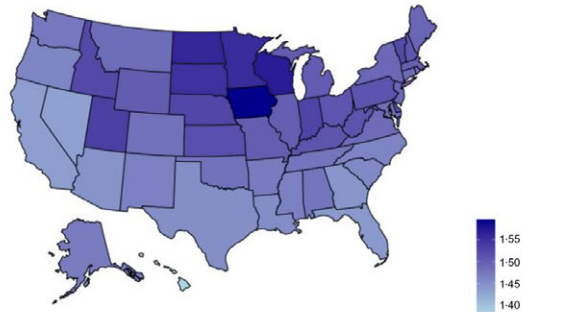
Mean daily intake of processed meat



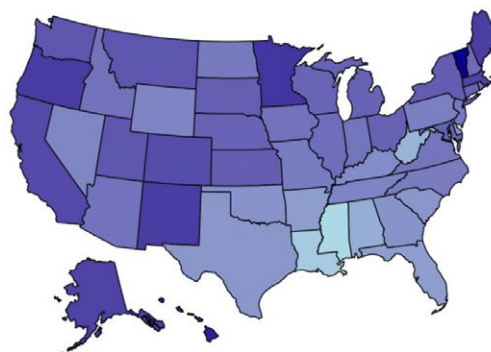
Mean vegetable intake (cups)



Mean dairy intake (cups)



Mean whole grain intake (ounces)



Mean daily intake of pizza

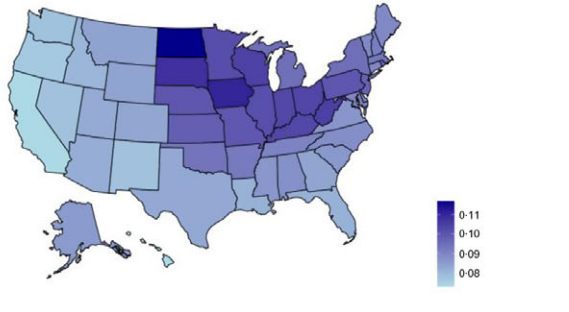


Fig. 4 Maps of average BMI and average intake of select dietary factors by state, 2017–2019

Table 2 Survey-weighted mean (SE) intake of fruit, vegetables, whole grains, dairy, added sugars and sugar-sweetened beverages (SSB), by sample demographic characteristics

		Cups/ d of fruit	SE	Cups/d of vegetables	SE	Ounces/ d of whole grains	SE	Cups/ d of dairy	SE	Mean tsp/d added sugars	SE	Proportion drinking at least 1 sugar- sweetened beverage/d
Sex	Overall	0.83	0.0007	1.54	0.0008	0.56	0.0004	1.54	0.001	15.4	0.01	34 %
	Female	0.81	0.0006	1.43	0.0008	0.54	0.0004	1.36	0.0008	14.1	0.01	33 %
	Male	0.85	0.001	1.65	0.001	0.58	0.0008	1.72	0.002	16.7	0.02	35 %
Age	20–64	0.83	0.0008	1.53	0.0009	0.56	0.0004	1.55	0.001	15.7	0.01	35 %
	65–85	0.85	0.001	1.60	0.001	0.57	0.0008	1.49	0.002	13.8	0.02	28 %
Race	White	0.83	0.0007	1.55	0.0008	0.56	0.0004	1.55	0.001	15.4	0.01	33 %
	Black	0.85	0.006	1.48	0.006	0.56	0.004	1.38	0.008	15.5	0.1	38 %
	American Indian	0.80	0.01	1.47	0.01	0.56	0.007	1.49	0.016	16.2	0.2	46 %
	Asian	0.83	0.005	1.50	0.004	0.56	0.003	1.40	0.006	14.7	0.08	26 %
	Oceanian	0.81	0.03	1.48	0.03	0.52	0.01	1.46	0.03	15.0	0.4	32 %
	Multiple	0.83	0.002	1.53	0.002	0.56	0.001	1.53	0.004	15.7	0.04	37 %
Ethnicity	Non-Hispanic	0.83	0.0007	1.54	0.0008	0.56	0.0004	1.54	0.001	15.4	0.01	34 %
	Hispanic	0.83	0.002	1.52	0.003	0.59	0.0005	1.54	0.004	15.4	0.05	36 %
Education	Less than college	0.81	0.0009	1.50	0.001	0.55	0.0006	1.54	0.001	15.8	0.02	38 %
	College or more	0.89	0.0008	1.63	0.001	0.59	0.0005	1.54	0.001	14.5	0.01	26 %
BMI	Underweight	0.87	0.007	1.45	0.007	0.56	0.004	1.48	0.01	16.1	0.1	37 %
	Normal weight	0.88	0.001	1.56	0.001	0.58	0.0008	1.48	0.002	15.0	0.02	31 %
	Overweight	0.84	0.001	1.58	0.001	0.56	0.0008	1.55	0.002	15.3	0.02	33 %
	Obese	0.78	0.001	1.49	0.001	0.54	0.0007	1.57	0.002	15.8	0.02	36 %

Table 3 Percent of the survey weighted 23andMe research participant population meeting federal dietary recommendations for fruits, vegetables and dairy by age and sex (2017–2019)

		Fruit	Vegetables	Dairy
Male	Overall	2.6	5.9	2.8
	20–30	4.2	1.1	7.7
	31–50	1.6	1.6	5.5
Female	51+	2.3	6.1	3.4
	20–30	0.4	0.3	1.1
	31–50	3.7	8.4	0.9
	51+	3.0	10.8	0.6

juice will be systematically misclassified by the DSQ as not meeting fruit intake recommendations. In an assessment of this screener, it was recommended to pilot test in each population prior to use, and to take caution when deriving precise estimates, but neither ceiling effects on fruit intake nor this particular problem was noted⁽¹¹⁾. Because the 23andMe respondent population includes an over representation of college educated people who we show here consume less added sugars compared with those without a college degree, limitations inherent to this questionnaire may be more pronounced.

Generalisability of the survey-weighted 23andMe respondent sample to the US population was achievable for age, sex, education, white *v.* non-white and BMI. However, limits to sample weighting still require a higher representation of non-white participants to develop a generalisable weighted sample. An alternative approach would be sub-sampling the database⁽²⁹⁾ to achieve a more

balanced distribution across race/ethnicity, which we would recommend for analyses which aim to draw direct comparisons to the general population from the 23andMe research participant database. In addition to considerations of generalisability, other limitations to our study include the cross-sectional nature of the ascertainment for both exposure (diet) and outcome (BMI), which limits inferences of temporality and causality. Future studies may use prospectively collected data or Mendelian randomisation analyses to further understand the causal architecture of our observed associations. Finally, our results are based on a one-time dietary measurement, which may not reflect habitual dietary intake.

In addition to the very low proportion of participants who met national guidelines for intake of fruits and vegetables, plots of reported intake over time are not encouraging. While some patterns are easily explained by broader trends, for example, with milk, declines in reported consumption of fruits and vegetables over the study period are either indicative of changing characteristics of the 23andMe research population or are reflective of broader declines in consumption of these foods.

Our exploration of nutrition patterns within the 23andMe database identified several unique advantages of this large-scale, participant-driven, digital cohort. The ability to quickly and contemporaneously collect nutritional information can inform more expeditious assessment of nutrition trends compared with traditional surveys; in contrast to the 23andMe data, NHANES data are usually made available to researchers several years after collection. Additionally,



Table 4 Survey-weighted mean (SE) intake of select vegetables, fruit, dairy, added sugars and whole grains by region (2017–2019)

Region	n weighted	Mean cups/d of fruit	SE	Mean cups/d of vegetables	SE	Mean ounces/d whole grains	SE	Mean cups/d of dairy	SE	Mean tsp/d added sugars	SE
Midwest	93 803	0.83	0.002	1.49	0.002	0.56	0.001	1.59	0.003	15.9	0.03
Northeast	72 416	0.86	0.002	1.55	0.002	0.56	0.001	1.56	0.003	15.2	0.03
South	162 572	0.80	0.001	1.53	0.001	0.55	0.0007	1.52	0.002	15.7	0.03
West	152 792	0.85	0.001	1.58	0.001	0.57	0.0008	1.51	0.002	14.9	0.02

Food/Drink Item	Beta	Lower CI	Upper CI
Processed meat	0.215	0.212	0.219
Red meat	0.198	0.194	0.202
Fried potatoes	0.191	0.187	0.195
Pastry	0.156	0.153	0.160
Cheese	0.130	0.127	0.134
Pizza	0.122	0.118	0.126
Ice cream	0.101	0.097	0.104
Cookies/Cakes	0.066	0.062	0.070
Popcorn	0.064	0.061	0.068
Eggs	0.047	0.042	0.052
Other potato	0.047	0.043	0.051
Tomato sauce	0.034	0.030	0.038
Chocolate	0.002	-0.002	0.006
Salsa	-0.003	-0.007	0.0008
Whole grain bread	-0.021	-0.025	-0.017
Beans	-0.036	-0.039	-0.032
Cereal	-0.040	-0.044	-0.037
Other whole grains	-0.123	-0.126	-0.119
Other vegetables	-0.15	-0.154	-0.146
Salad	-0.163	-0.166	-0.159
Whole fruit	-0.171	-0.174	-0.167

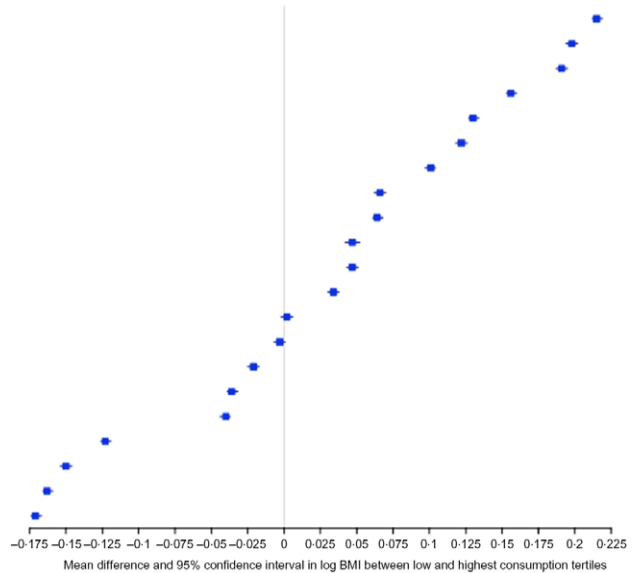


Fig. 5 Mean difference in log BMI (beta, 95 % CI) between the highest and lowest consumption tertiles. Beta estimates and 95 % CI were derived from linear models regressing intake tertile on log-transformed BMI, adjusted for race/ethnicity, education, age (centred), sex and age (centred) squared

our unprecedentedly large sample size enables well-powered subgroup analyses, including future nutrigenomic studies, and rigorous ascertainment of spatial and seasonal variation in dietary intake. Finally, digital ascertainment of nutritional exposures minimises participant burden, encouraging participation in future surveys⁽³⁰⁾.

In conclusion, we have characterised the demographic, seasonal and spatial patterns of nutritional habits among 23andMe research participants in the USA. Additionally, we report cross-sectional positive associations between BMI and the intake of red and processed meat as well as dairy and inverse associations between BMI and the intake of fruit, vegetables and whole grains. Our dataset offers a unique opportunity for rapid, large-scale, real-time data collection, which can inform national trends in a much shorter time frame than current nationwide surveys. Efforts to diversify the 23andMe research participant database will increase the generalisability to the US population, but because of the significant number of participants, survey sampling and weighting methods can achieve this at present. While large-scale cohorts like the 23andMe participants offer exciting future opportunities in precision

nutrition, general efforts to continuously work towards improvements in dietary habits remain critical for maximising health, maintaining a healthy BMI and preventing chronic diseases.

Acknowledgements

Acknowledgements: The authors would like to thank the research participants and employees of 23andMe for making this work possible. Members of the 23andMe Research Team are Michelle Agee, Adam Auton, Robert K. Bell, Katarzyna Bryc, Sarah K. Clark, Sarah L. Elson, Kipper Fletez-Brant, Pierre Fontanillas, Nicholas A. Furlotte, Pooja M. Gandhi, Karl Heilbron, Barry Hicks, David A. Hinds, Karen E. Huber, Ethan M. Jewett, Yunxuan Jiang, Aaron Kleinman, Keng-Han Lin, Nadia K. Litterman, Jennifer C. McCreight, Matthew H. McIntyre, Kimberly F. McManus, Joanna L. Mountain, Sahar V. Mozaffari, Priyanka Nandakumar, Elizabeth S. Noblin, Carrie A.M. Northover, Jared O’Connell, Steven J. Pitts, G. David



Poznik, J. Fah Sathirapongsasuti, Anjali J. Shastri, Suyash Shringarpure, Chao Tian, Joyce Y. Tung, Robert J. Tunney, Vladimir Vacic and Xin Wang. *Financial support*: This study was conducted with no external financial support. *Conflict of interest*: All authors are employees of 23andMe, Inc., and hold stock or stock options in 23andMe. All authors had full access to all of the data in this study and take complete responsibility for the integrity of the data and the accuracy of the data analysis. *Authorship*: J.S., R.G. and S.A. formulated the research question. B.C. carried out statistical analyses. J.S. and S.A. wrote the article. *Ethics of human subject participation*: This study was conducted according to the guidelines laid down in the Declaration of Helsinki, and all procedures involving research study participants were approved by the Ethical & Independent Review Services, a private independent institutional review board. All 23andMe customers consent to participate in research online and consent is captured electronically. Ethical & Independent Review Services approved this form of informed consent and waived the requirement to obtain signed consent under 45 CFR 46.117(c).

Supplementary material

To view supplementary material for this article, please visit <https://doi.org/10.1017/S1368980020001251>

References

- GBD 2016 Risk Factors Collaborators (2017) Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* **390**, 1345–1422.
- The US Burden of Disease Collaborators (2018) The State of US Health, 1990–2016: burden of diseases, injuries, and risk factors among US states. *JAMA* **319**, 1444–1472.
- Micha R, Penhalvo JL, Cudhea F *et al.* (2017) Association between dietary factors and mortality from heart disease, stroke, and type 2 diabetes in the United States. *JAMA* **317**, 912–924.
- Rehm CD, Penhalvo JL, Afshin A *et al.* (2016) Dietary intake among US adults, 1999–2012. *JAMA* **315**, 2542–2553.
- Willett WC, Koplan JP, Nugent R *et al.* (2006) Prevention of chronic disease by means of diet and lifestyle changes. In *Disease Control Priorities in Developing Countries*, 2nd ed. [DT Jamison, JG Breman, AR Measham *et al.*, editors]. New York: Oxford University Press.
- Global action plan for the prevention and control of NCDs 2013–2020. World Health Organization. <https://www.who.int/nmh/publications/ncd-action-plan/en/> (accessed October 2019).
- Corella D, Peloso G, Arnett DK *et al.* (2009) APOA2, dietary fat, and body mass index: replication of a gene-diet interaction in 3 independent populations. *Arch Intern Med* **169**, 1897–1906.
- Hu FB, Willett WC (2018) Current and future landscape of nutritional epidemiologic research. *JAMA* **320**, 2073–2074.
- NHANES Dietary Screener Questionnaire, Self-Administered Questionnaire. Web. <https://epi.grants.cancer.gov/nhanes/dietscreen/questionnaires.html#web> (accessed April 2019).
- National Cancer Institute. *Dietary Screener Questionnaire (DSQ) in the NHANES 2009–10: Relationship between Dietary Factors & Food Items on the DSQ*. <https://epi.grants.cancer.gov/nhanes/dietscreen/relationship.html> (accessed September 2019).
- Thompson FE, Midthune D, Kahle L *et al.* (2017) Development and evaluation of the national cancer institute's dietary screener questionnaire scoring algorithms. *J Nutr* **147**, 1226–1233.
- National Cancer Institute. *Data Processing & Scoring Procedures Using Current Methods (Recommended)*. <https://epi.grants.cancer.gov/nhanes/dietscreen/scoring/current/> (accessed October 2019).
- U.S. Department of Health and Human Services and U.S. Department of Agriculture (2015) *2015–2020 Dietary Guidelines for Americans*. 8th ed. <https://health.gov/dietaryguidelines/2015/guidelines/> (accessed October 2019).
- US Census Bureau (2019) *Quick Facts United States*. <https://www.census.gov/quickfacts/fact/table/US/PST045217> (accessed April 2019).
- Centers for Disease Control and Prevention. National Center for Health Statistics (2016) *Table 058: Normal weight, overweight, and obesity among adults aged 20 and over, by selected characteristics: United States, selected years 1988–1994 through 2011–2014*. <https://www.cdc.gov/nchs/data/hus/2016/058.pdf> (accessed April 2019).
- Lumley T (2018) Survey: analysis of complex survey samples. R package version 3.29.
- Battaglia MP, Izrael D, Hoaglin DC *et al.* (2004) Tips and tricks for raking survey data (a.k.a. sample balancing). Presented at the 57th Annual Meeting of the World Association for Public Opinion Research, Phoenix, AZ, 11 May 2004.
- Bentley J & Kantor L (2019) *Food Availability (Per Capita) Data System*. Washington, DC: USDA, Economic Research Service.
- Rehm CD, Penhalvo JL, Afshin A *et al.* (2016) Dietary intake among US adults, 1999–2012. *JAMA* **315**, 2542–2553. doi: 10.1001/jama.2016.7491.
- United States Department of Agriculture. All About the Fruit Group. <https://www.choosemyplate.gov/eathealthy/fruit> (accessed April 2019).
- United States Department of Agriculture. *All About the Vegetable Group*. <https://www.choosemyplate.gov/eathealthy/vegetables> (accessed April 2019).
- Centers for Disease Control and Prevention (2002) *National Health and Nutrition Examination Survey. MEC In-Person Dietary Interviewers Procedures Manual*. https://www.cdc.gov/nchs/data/nhanes/nhanes_03_04/dietary_mec.pdf (accessed April 2019).
- Behavioral Risk Factor Surveillance System Published Maps*. <https://www.cdc.gov/nutrition/data-statistics/sugar-sweetened-beverages-intake.html> (2017) and obesity <https://www.cdc.gov/obesity/data/prevalence-maps.html> (2019) (accessed October 2019).
- Centers for Disease Control and Prevention (2013) *Prevalence of Sugar-Sweetened Beverage Intake Among Adults – 23 States and the District of Columbia*. <https://www.cdc.gov/mmwr/volumes/65/wr/mm6507a1.htm> (accessed April 2019).
- Ioannidis JP (2013) Implausible results in human nutrition research. *BMJ* **347**, f6698.
- Newby PK, Muller D, Hallfrisch J *et al.* (2003) Dietary patterns and changes in body mass index and waist circumference in adults. *Am J Clin Nutr* **77**, 1417–1425.
- Hoerger M (2010) Participant dropout as a function of survey length in internet-mediated university studies: implications



- for study design and voluntary participation in psychological research. *Cyberpsychol Behav Soc Netw* **13**, 697–700.
28. Walker RW, Dumke KA & Goran MI (2014) Fructose content in popular beverages made with and without high-fructose corn syrup. *Nutrition* **30**, 928–935.
 29. Liu H, Cella D, Gershon R *et al.* (2010) Representativeness of the patient-reported outcomes measurement information system internet panel. *J Clin Epidemiol* **63**, 1169–1178.
 30. Bowling A (2005) Mode of questionnaire administration can have serious effects on data quality. *J Public Health* **27** 281–291.