Received: 12 September 2018 | Revised: 30 October 2018 | Accepted: 21 November 2018

DOI: 10.1002/bdr2.1441

WILEY Birth Defects Research

ORIGINAL ARTICLE

A machine learning approach to investigate potential risk factors for gastroschisis in California

Kari A. Weber[1] | Wei Yang[1] | Suzan L. Carmichael[1] | Amy M. Padula[2] | Gary M. Shaw[1]

[1]Department of Pediatrics, Division of Neonatology, Stanford University School of Medicine, Stanford, California

[2]Department of Obstetrics, Gynecology and Reproductive Sciences, University of California, San Francisco, California

Correspondence
Kari A. Weber, Department of Pediatrics, Stanford University School of Medicine, 1265 Welch Road x1C21, Stanford, California 94305.
Email: kaweber@stanford.edu

Funding information
Eunice Kennedy Shriver National Institute of Child Health and Human Development, Grant/Award Number: R01HD075761; National Institute of Environmental Health Sciences, Grant/Award Number: P01ES022849, R00ES021470; University of North Carolina Department of Nutrition Clinical Research Center, Nutrition Epidemiology Core, Grant/Award Number: DK56350; Centers for Disease Control and Prevention Centers of Excellence, Grant/Award Number: U01DD001033

Background: To generate new leads about risk factors for gastroschisis, a birth defect that has been increasing in prevalence over time, we performed an untargeted data mining statistical approach.

Methods: Using data exclusively from the California Center of the National Birth Defects Prevention Study, we compared 286 cases of gastroschisis and 1,263 nonmalformed, live-born controls. All infants had delivery dates between October 1997 and December 2011 and were stratified by maternal age at birth (<20 and ≥ 20 years). Cases and controls were compared by maternal responses to 183 questions (219 variables) using random forest, a data mining procedure. Variables deemed important by random forest were included in logistic regression models to estimate odds ratios and 95% confidence intervals.

Results: Among women younger than 20, of variables deemed important, there were higher odds observed for higher consumption of chocolate, low intake of iron, acetaminophen use, and urinary tract infections during the beginning of pregnancy. After adjustment, the higher odds remained for low iron intake and a urinary tract infection in the first month of pregnancy. Among women aged 20 or older, of variables deemed important, higher odds were observed for US-born women of Hispanic ethnicity and for parental substance abuse. There were lower odds observed for obese women, women who ate any cereal the month before pregnancy, and those with higher parity.

Conclusions: We did not discover many previously unreported associations, despite our novel approach to generate new hypotheses. However, our results do add evidence to some previously proposed risk factors.

KEYWORDS

data mining, etiology, gastroschisis, maternal age, random forest, teenage pregnancy

1 | INTRODUCTION

Gastroschisis is a congenital anomaly of the abdominal wall where part of the intestines and sometimes other internal organs are outside the body at birth. Its prevalence in the US is roughly four per 10,000 live births (Canfield et al., 2006; Parker et al., 2010). The prevalence has been increasing over the past several decades (Castilla, Mastroiacovo, & Orioli, 2008; Jones et al., 2016; St Louis et al., 2017; Vu, Nobuhara, Laurent, & Shaw, 2008). Many factors have been investigated one, or a few at a time, and no study has sufficiently explained the increase in temporal prevalence or clearly identified a risk factor for gastroschisis other than the well described increased (fivefold) risk to women 20 years of age and younger (Gill et al., 2012; Rasmussen & Frias, 2008; Reefhuis & Honein, 2004; Werler, Mitchell, & Shapiro, 1992). Further, even fewer investigations have been made to specifically disentangle the potential disproportionate risks

factors that may underlie prevalence differences between younger (<20 years) and older (≥20 years) mothers. Given a general lack of information on the potential multiple-factor-influences on gastroschisis risk, as well as a need to generate new leads about risk factors, we performed an untargeted statistical approach to fully utilize the rich set of available data in California.

2 | METHODS

2.1 | Study population and data collection

Data are from the California Center of the National Birth Defects Prevention Study (NBDPS) (Reefhuis et al., 2015). The California Center is a joint effort between Stanford University and the California Birth Defects Monitoring Program in the Department of Public Health. The California Birth Defects Monitoring Program has been performing population-based active surveillance and collecting data from women residing in one of eight counties in the San Joaquin Valley at time of delivery since 1986 (Croen, Shaw, Jensvold, & Harris, 1991). To ascertain cases, staff visited hospitals with obstetric or pediatric services, clinical genetics prenatal and postnatal outpatient services, and cytogenetic laboratories.

Included cases of gastroschisis were confirmed by a clinical geneticist using clinical, surgical, or autopsy reports. Cases suspected of having a chromosomal abnormality or identifiable syndrome were ineligible (Rasmussen et al., 2003). Controls were non-malformed, live-born infants randomly selected from birth hospitals of cases. Roughly 150 controls were selected per same study year as cases. Interviews with maternal participants took place between 6 weeks and 24 months after the estimated date of delivery and were performed using standardized, computer-based questionnaires in English or Spanish. Questions sought information on maternal health, pregnancy history, as well as maternal and paternal sociodemographic and behavioral information at specific points before and throughout pregnancy. Dietary information was collected via a 58-item food frequency questionnaire, developed and validated by Willet and colleagues (Willett et al., 1985), of average intake for the year before pregnancy and certain other specific intake questions. Time-periods were delineated by month as before (B3–B1) or during pregnancy (P1–P9). For this specific analysis, the time-periods of interest were the month before conception (B1) until the second month of pregnancy (P1–P2) as this is believed to be the relevant etiologic window.

The dataset included cases and controls with estimated dates of delivery from October 1, 1997 to December 31, 2011. Mothers were interviewed for 286 cases and 1,263 controls. Participation in the interview was 66% for case and 64% for control mothers. Maternal participants missing >10% of questions were excluded to remove participants who did not complete the questionnaire, resulting in 268 cases and 1,203 controls. Cases and controls were compared for maternal responses to 183 questions (219 variables). This dataset was analyzed by age stratified as younger than 20 (109 cases and 181 controls) or 20 and older (159 cases and 1,022 controls).

A supplemental dataset also included pesticide and air pollution data for a portion of the cases and controls, that is, subjects with estimated dates of delivery from October 1, 1997 to December 31, 2006. Pesticide and air pollution exposure and associations with gastroschisis have been investigated previously and methods are described elsewhere (Padula et al., 2013; Shaw et al., 2014). Briefly, maternal residences were geocoded and linked with the California Department of Pesticide Regulation reporting records and US Environmental Protection Agency's Air Quality System database (https://aqs.epa.gov/aqsweb/documents/data_mart_welcome.html). Only those living in one of the eight counties for 75% of the time during B1–P2 were eligible for pesticide exposure assignment and during P1–P2 for air pollution exposure assignment. Pesticide data were collected for 461 chemicals and 69 physiochemical groupings that were applied at >100 lb and deemed toxic based on Environmental Protection Agency's (EPA) risk assessment and California Proposition 65 or were classified as endocrine disruptors. Data were collected for daily 24-hr averages of nitrogen oxide, nitrogen dioxide, particulate matter <10 μm (PM_{10}), particulate matter <2.5 μm ($PM_{2.5}$), and carbon monoxide, and a daily 8-hr maximum of ozone and averaged over the study period. Traffic-density measures were also collected based on distance-decayed annual average daily traffic volumes from the Geographic Data Technology traffic count data. Of participants with data on both pesticides and pollutants, missing <10% of questions, there were 145 cases and 758 controls. Cases and controls were compared for responses to 242 questions (278 variables).

2.2 | Statistical analysis

Cases and controls were compared for eligible variables using random forest. The variables are described in Appendix A. Random forest is a well-established data mining procedure that calculates a set of decision trees using random subsets of the data and combines them to produce a mean prediction model of case status based on variable importance (Strobl, Malley, & Tutz, 2009). Random forest accounts for interactions and nonlinear associations among a large group of factors simultaneously to determine the importance of individual variables (Strobl et al., 2009). The current version of the program removes a previously existing bias toward correlated variables by using conditional inference trees (Strobl et al., 2009).

We calculated a variable importance measurement for each potential predictor variable, using the "varimp" function in the party package in R software to obtain the metric

mean decrease accuracy (MDA) specifying "ntree = 2,500" and "mtry = 15" as the number of trees and number of selected predictor variables per split, respectively. If the MDA value was above the absolute value of the MDA for the lowest negative MDA, variables were considered "important." The importance of irrelevant variables should vary randomly around zero (Strobl et al., 2009).

Variables were included in the random forest procedure if questions occurred with a frequency ≥ 0.1% to increase the likelihood of having adequate exposed participants for calculation. Time-varying variables were included for the time-period from 1 month before pregnancy (B1) until the second month of pregnancy (P1–P2) to reflect the hypothesized etiologically relevant window (Lammer, Iovannisci, Tom, Schultz, & Shaw, 2008). Pesticide exposure was included as any exposure during B1–P2 and air pollution values were included for the time-period P1–P2 based on previously collected data (Padula et al., 2013; Shaw et al., 2014).

Nutrient and air pollution values were divided into three categories based on values among controls: those with exposure levels <25%, those within the interquartile range, and those with exposure levels >75%. For the initial random forest, simple imputation was performed for missing data. For the nutrient and air pollution values, all missing data for these variables were imputed as—being within the interquartile range. Other missing data were imputed with the median among controls for continuous variables and the most frequently occurring response among controls for categorical variables.

The random forest analysis only provided the ranking list of "important" variables, thus the effect of these variables toward the outcomes in magnitude and direction were further examined using traditional parametric models. For the parametric models, multiple imputation was performed using the MICE package in R (R Core Team, 2013) to impute missing data based on all other variables to create 20 imputed datasets. Random forest results differ based on the dataset so to be robust, variables deemed to be in the top five important variables by random forest in at least one of these imputed datasets were included in a multivariable logistic regression model where pooled adjusted odds ratios (ORs) and 95% confidence intervals (95% CIs) were estimated. ORs were adjusted for all other important variables.

Random forest accounts for correlation but logistic regression does not. After preliminary analyses, acetaminophen use in P1 and P2 in younger mothers were highly correlated ($r = .77$), as were reports of paternal substance abuse across time-periods in older mothers (all $r > .88$), thus we combined the respective variables in the logistic regression analyses. The resulting variables were any versus no acetaminophen use during P1–P2 and any versus no paternal substance abuse during B1–P2. A sensitivity analysis was also performed additionally adjusting for maternal age (as a continuous variable) to assess residual confounding.

Random forest analyses and multiple imputation were performed in R software using the Party Package and MICE Package, respectively (version 3.4.4) (R Core Team, 2013). All other analyses were performed using SAS version 9.4 (SAS Institute, Cary, NC).

3 | RESULTS

Employing random forest analyses for case and control mothers younger than age 20 (Figure 1) and among case and control mothers age 20 or older (Figure 2) we identified important predictors of gastroschisis including various dietary, demographic, and behavioral factors. Additionally, random forest was performed on the supplemental dataset for whom data were available on air pollution and pesticide exposures. The addition of these variables did not reveal further insights, that is, none of the air pollutants nor pesticide exposures emerged as important predictors. Because these environmental exposure variables did not add information, subsequent analyses did not include these variables.

Random forest performed on each of the 20 imputed datasets yielded mostly similar results. Analyses were performed separately for each age stratification and the variables appearing in the top five important predictors at least once were included in subsequent analyses. The 10 variables presenting as important predictors among women younger than age 20 were intake of chocolate, organ meats, apples or pears, and candy, iron, acetaminophen use during P1 and P2, having a urinary tract infection during P1, the number of people supported by the household income, and minutes spent bathing. The nine variables presenting as important among women age 20 or older were maternal race/ethnicity, maternal substance abuse during B1, paternal substance abuse during B1, P1, and P2, prepregnancy obesity, cereal intake during B1, paternal age, and parity.

Among women younger than 20 years of age (Table 1), higher odds of delivering infants with gastroschisis were observed for more frequent consumption of chocolate and moderate consumption of candy compared to the lowest consumption, being in the lowest 25% of iron intake compared the middle 50%, acetaminophen use during the first 2 months of pregnancy and having a urinary tract infection in the first month of pregnancy. After adjustment for other variables, patterns remained the same, although the associations were no longer statistically significant for consumption of chocolate and candy and acetaminophen use. There were inverse (and statistically imprecise) associations observed for consumption of organ meats 1–3x per month and the highest consumption of apples or pears.

Among women age 20 or older (Table 2), higher odds of delivering infants with gastroschisis were observed for US-born women of Hispanic ethnicity or those identifying as Other race/ethnicity compared to white non-Hispanic women. Higher odds were also observed for maternal

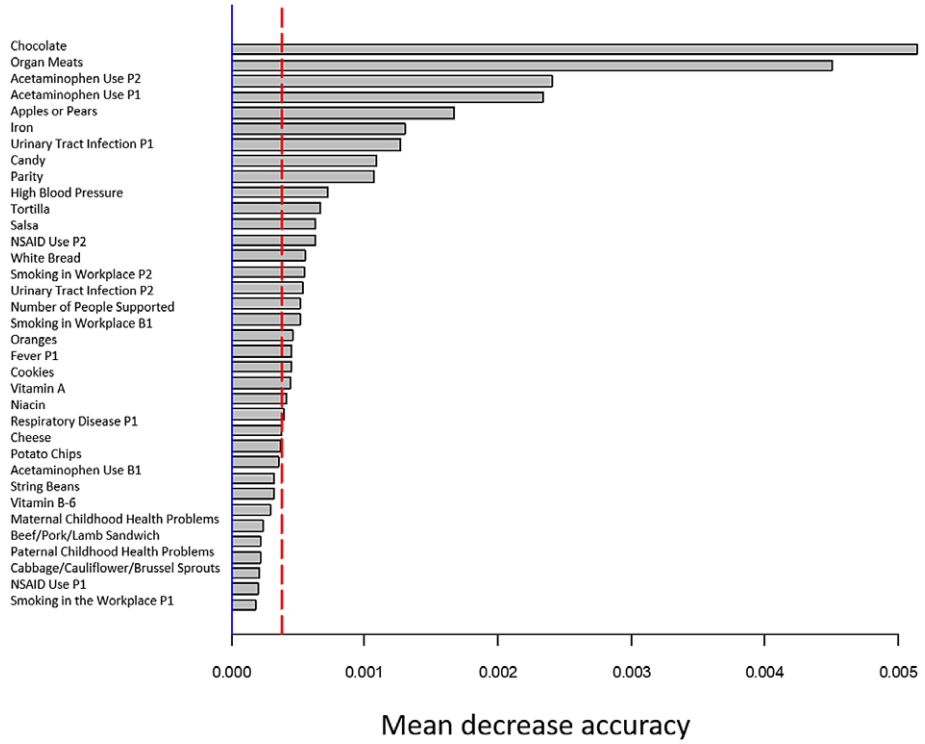

FIGURE 1 Random forest ranking plot of variable importance in predicting risk of gastroschisis for 109 cases and 181 controls born to women less than age twenty. Predictors to the right of the dashed vertical line are important. Nutrient values were divided into three categories based on values among controls: Those with exposure levels <25%, those within the interquartile range, and those with exposure levels >75%. All missing data for these variables were imputed as—Being within the interquartile range. Other missing data were imputed with the median among controls for continuous variables and the most frequently occurring response among controls for categorical variables. Categorical variables were (yes/no) except food frequency questionnaire items are divided into the following categories: <1x per month, 1–3x per month, 1x per week, 2-6x per week, or > 6x per week. Continuous variables included parity and number of people supported

substance abuse in the month before pregnancy and paternal substance abuse any time during the study period. Inverse associations with gastroschisis were observed for women who were obese prepregnancy compared to those who were not, cereal intake in the month before pregnancy, and parity. Paternal age was also inversely associated with gastroschisis among older mothers. However, parental ages were correlated and thus it is probable that the observed association with (younger) paternal age was a proxy for the residual risk associated with younger maternal age (e.g., 20–25 years) within this stratified group of women ≥20. In the additional sensitivity analysis performed, the addition of maternal age did not alter any of the results for women <20 but among women ≥20, maternal age was inversely associated with gastroschisis (aOR 0.88, 95% CI 0.83–0.94) and associations with paternal age and parity were attenuated and no longer significant. The addition of maternal age did not alter any other results (results not shown).

4 | DISCUSSION

This analysis utilized a well-established, untargeted statistical approach, random forest, in an effort to find new risk factors for gastroschisis. We explored potential differences in risk factors between younger and older mothers by simultaneously evaluating a large number of variables available in California separately in these groups. We observed different factors to be important predictors in these two populations and thus observed different statistically significant associations. Among women younger than age 20, important predictors associated with gastroschisis were consumption of certain foods, acetaminophen use, and infection, whereas among older women, important predictors associated with gastroschisis were certain maternal race/ethnicities, parental substance abuse, obesity, and paternal age. Exposures to pesticides or various air pollutants did not contribute as important predictors to either maternal age group.

Among women younger than 20, a few of the important variables were nutrients and food intake items. We observed a positive association between low iron intake and delivering an infant with gastroschisis as well as for moderate consumption of candy. The same pattern was observed for moderate amounts of chocolate intake but a suggestive inverse association was observed for intake >6 times per week and for high intake of apples or pears, sources of vitamins such as vitamin C. Previous studies have observed higher odds of gastroschisis with low alpha-carotene intake, low glutathione intake, and high nitrosamine intake after adjustment for other factors (Torfs, Lam, Schaffer, & Brand, 1998) and suggestive inverse associations with higher intake of protein, fat,

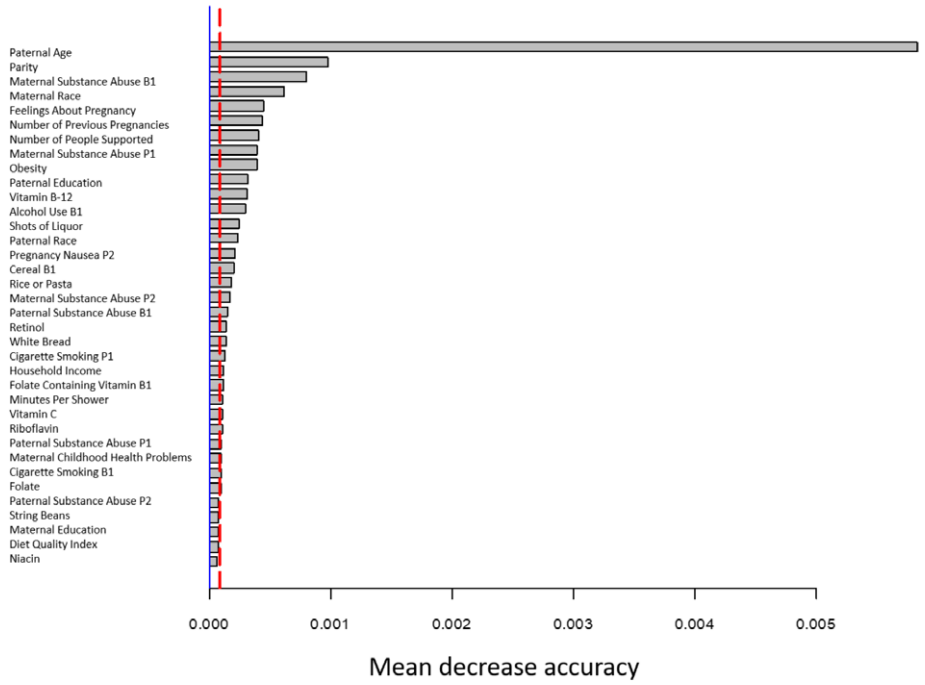

FIGURE 2 Random forest ranking plot of variable importance in predicting risk of gastroschisis for 159 cases and 1,022 controls born to women age twenty or older. Predictors to the right of the dashed vertical line are important. Nutrient values were divided into three categories based on values among controls: Those with exposure levels <25%, those within the interquartile range, and those with exposure levels >75%. All missing data for these variables were imputed as—Being within the interquartile range. Other missing data were imputed with the median among controls for continuous variables and the most frequently occurring response among controls for categorical variables. Categorical variables were (yes/no) except maternal race/ethnicity (white non-Hispanic, Hispanic foreign-born, Hispanic US-born, black non-Hispanic, other), feelings about pregnancy (wanted to be pregnant, wanted to wait until later, did not care, pregnant despite consistent contraceptive use), household income (<$10,000, $10,000–$50,000, >$50,000) and food frequency questionnaire items are divided into the following categories: <1x per month, 1–3x per month, 1x per week, 2–6x per week, or > 6x per week. Continuous variables included paternal age, parity, number of previous pregnancies, number of people supported and minutes per shower

alpha-carotene, and magnesium and inverse associations for higher intake of copper, folate, vitamin B12, and oleic acid (Feldkamp et al., 2011). Another study observed significant inverse associations for both Diet Quality Index and Mediterranean diet score only among Hispanic women (Feldkamp, Krikov, Botto, Shaw, & Carmichael, 2014). In the current study, we observed inverse associations with slight increases in organ meat intake among women younger than 20, a source of vitamin A. We also observed an inverse association with cereal intake in the month before pregnancy among women age 20 or older, a food containing folic acid. The collection of these observations indicates that risk among younger women is influenced by aspects of the diet or behaviors related to aspects of the diet. Deeper inquiry into these domains seem warranted.

We observed higher odds of delivering an infant with gastroschisis with acetaminophen use in the beginning of pregnancy among women younger than 20 but after adjustment for other variables such as urinary tract infection the odds were attenuated. Associations between maternal use of various medications and gastroschisis have been studied previously (Ahrens et al., 2013; Alwan, Reefhuis, Rasmussen, Olney, & Friedman, 2007; Draper et al., 2008; Feldkamp, Meyer, Krikov, & Botto, 2010; Interrante et al., 2017; Lin et al., 2008; Polen, Rasmussen, Riehle-Colarusso, & Reefhuis, 2013; Waller et al., 2010; Werler et al., 1992; Werler, Sheehan, & Mitchell, 2002). However, acetaminophen was the only important predictor we identified of the medications included in our analysis. A previous analysis in NBDPS observed an inverse association between acetaminophen use and gastroschisis among women who also had an infection and fever (Feldkamp et al., 2010). It is unclear whether the medication or the indication for medication is the risk factor in many studies of medication use. A previous study also observed an association between having a urinary tract infection and gastroschisis among younger mothers (Yazdy, Mitchell, & Werler, 2014) as we did, and another study observed a positive association between having a fever and gastroschisis (Waller et al., 2018) suggesting that further research into the joint effects of infection, medication use and indication, and maternal age may be warranted.

Another factor only found to be important among women age 20 or older was maternal race/ethnicity. We observed increased odds of having an infant with gastroschisis for women identifying as Hispanic US-born or other race/ethnicity compared to White Non-Hispanic women. This is consistent with a study that observed higher odds in Hispanic US-born women only among older mothers (Khodr et al., 2013) but not with an age-matched study in California (Torfs, Velie, Oechsli, Bateson, & Curry, 1994).

TABLE 1 Pooled[a] summary statistics (number and percent for categorical, mean and standard deviation for continuous) and odds ratios for Important[b] predictors of Gastroschisis in infants born to women age < 20 years, California, 1997–2011

Variables	Cases (n = 109) No. (%)	Controls (n = 181) No. (%)	OR	(95% CI)	aOR [c]	(95% CI)
Intake of chocolate (1 oz.)						
<1x per month	22 (20.2)	51 (28.2)	Reference		Reference	
1–3x per month	15 (13.8)	39 (21.5)	0.89	(0.41–1.94)	0.57	(0.22–1.51)
1x per week	23 (21.1)	25 (13.8)	2.13	(1.00–4.54)	2.16	(0.89–5.24)
2–6x per week	40 (36.7)	35 (19.3)	2.65	(1.35–5.20)	2.04	(0.88–4.74)
>6x per week	9 (8.3)	31 (17.1)	0.67	(0.28–1.65)	0.49	(0.17–1.42)
Intake of organ meats (3–4 oz.)						
<1x per month	79 (72.5)	105 (58.0)	Reference		Reference	
1–3x per month	15 (13.8)	54 (29.8)	0.37	(0.19–0.70)	0.31	(0.15–0.66)
1x per week	13 (11.9)	12 (6.6)	1.44	(0.62–3.33)	1.79	(0.65–4.92)
2–6x per week	2 (1.8)	10 (5.5)	NA		NA	
Intake of apples or pears (1)						
<1x per month	18 (16.5)	24 (13.3)	Reference		Reference	
1–3x per month	17 (15.6)	16 (8.8)	1.42	(0.57–3.57)	2.47	(0.76–8.01)
1x per week	22 (20.2)	21 (11.6)	1.40	(0.60–3.30)	1.36	(0.49–3.73)
2-6x per week	38 (34.9)	67 (37.0)	0.76	(0.37–1.57)	0.75	(0.30–1.83)
>6x per week	14 (12.8)	53 (29.3)	0.35	(0.15–0.83)	0.45	(0.16–1.22)
Intake of candy (1 oz.)						
<1x per month	25 (22.9)	62 (34.3)	Reference		Reference	
1–3x per month	16 (14.7)	38 (21.0)	1.04	(0.50–2.20)	1.07	(0.43–2.66)
1x per week	28 (25.7)	22 (12.2)	3.16	(1.53–6.52)	2.37	(1.00–5.61)
2–6x per week	26 (23.9)	36 (19.9)	1.79	(0.90–3.56)	1.61	(0.68–3.82)
>6x per week	14 (12.8)	23 (12.7)	1.51	(0.67–3.40)	2.18	(0.80–5.92)
Dietary intake of iron						
<25th percentile	42 (38.5)	42 (23.2)	1.96	(1.13–3.39)	2.19	(1.12–4.30)
25–50th percentile	49 (45.0)	96 (53.0)	Reference		Reference	
≥75th percentile	18 (16.5)	43 (23.8)	0.82	(0.43–1.57)	1.07	(0.46–2.49)
Acetaminophen use P1–P2	43 (39.4)	43 (23.8)	2.09	(1.25–3.50)	1.63	(0.87–3.04)
Urinary tract infection P1	11 (10.1)	4 (2.2)	4.97	(1.54–16.0)	4.87	(1.20–19.8)
	Mean (SD)	**Mean (SD)**	**OR**	**(95% CI)**	**aOR[c]**	**(95% CI)**
Number of people household income supports [d]	3.1 (2.3)	3.6 (2.0)	0.89	(0.78–1.01)	0.98	(0.84–1.13)
Minutes per bath [d]	13 (17.3)	18 (21.1)	0.99	(0.97–1.00)	0.98	(0.96–1.00)

[a] Results were combined and averaged across all 20 imputed datasets.
[b] Important variables were determined based on the "varimp" function in the random forest analyses and the resulting mean decrease accuracy metric. Variables deemed to be in the top five important variables in at least one of the imputed datasets were included.
[c] Adjusted for the variables included in the table.
[d] Variables had missing values among women age < 20 years old and were imputed.

The same California study did observe a higher odds of delivering an infant with gastroschisis for women with a lower income and a few other markers of lower socioeconomic status and gastroschisis (Torfs et al., 1994). Once multiple imputation was performed, the only marker of socioeconomic status in the top important predictors was number of people supported by the household income and only among women younger than 20. There was an inverse association between the number of people and gastroschisis, however, it is not immediately clear if a higher number of dependents corresponds to higher or lower status.

Among women age 20 or older, three additional factors that were important and associated with gastroschisis were substance abuse, lack of obesity, and parity. In our analyses, substance abuse combined use of marijuana, hash, cocaine, crack, hallucinogens, heroin, mushrooms, and other. Previous studies have observed positive associations between maternal use of marijuana (Torfs et al., 1994) or any recreational drug use (Draper et al., 2008) in the first trimester and gastroschisis. However, these studies did not explore paternal substance abuse. Both maternal and paternal substance abuse were associated with gastroschisis, although they were attenuated after adjustment, possibly due to correlation between the variables. Substance abuse could have biological effects on each parent or could be a proxy of other risky behaviors. Previous studies have also observed higher odds

TABLE 2 Pooled[a] summary statistics (number and percent for categorical, mean and standard deviation for continuous) and odds ratios for Important[b] predictors of Gastroschisis in infants born to women age ≥ 20 years, California, 1997–2011

Variables	Cases (n = 159) No. (%)	Controls (n = 1,022) No. (%)	OR	(95% CI)	aOR[c]	(95%CI)
Maternal race/ethnicity						
White non-Hispanic	35 (22.0)	305 (29.8)	Reference		Reference	
Hispanic foreign-born	34 (21.4)	359 (35.1)	0.82	(0.50–1.35)	1.06	(0.62–1.82)
Hispanic US-born	63 (39.6)	239 (23.4)	2.30	(1.47–3.59)	2.14	(1.31–3.49)
Black non-Hispanic	4 (2.5)	31 (3.0)	1.13	(0.38–3.40)	1.18	(0.36–3.95)
Other	23 (14.5)	88 (8.6)	2.27	(1.28–4.05)	2.27	(1.21–4.27)
Maternal substance abuse B1[d]	23 (14.5)	37 (3.6)	4.50	(2.60–7.81)	2.19	(1.07–4.49)
Paternal substance abuse B1–P2	43 (27.0)	98 (9.6)	3.42	(2.27–5.16)	1.77	(1.04–3.02)
Pre-pregnancy obesity (≥30 kg/m^2)	12 (7.5)	229 (22.4)	0.27	(0.14–0.51)	0.31	(0.16–0.60)
Cereal intake B1	107 (67.3)	796 (77.9)	0.59	(0.41–0.85)	0.57	(0.38–0.84)
	Mean (SD)	**Mean (SD)**	**OR**	**(95% CI)**	**aOR[c]**	**(95%CI)**
Paternal age (years)	25.4 (5.5)	29.8 (6.6)	0.87	(0.84–0.90)	0.90	(0.87–0.94)
Parity[d] (previous live births)	0.8 (1.1)	1.4 (1.4)	0.65	(0.55–0.77)	0.80	(0.67–0.96)

[a] Results were combined and averaged across all 20 imputed datasets.
[b] Important variables were determined based on the "varimp" function in the random forest analyses and the resulting mean decrease accuracy metric. Variables deemed to be in the top five important variables in at least one of the imputed datasets were included.
[c] Adjusted for the variables included in the table.
[d] No missing values were imputed for this variable among women age ≥ 20.

of gastroschisis among women who drank alcohol during the periconceptional period (Richardson et al., 2011). Our study observed and inverse association between prepregnancy obesity and gastroschisis and are consistent with previous findings (Lam, Torfs, & Brand, 1999; Waller et al., 2007). Additional research into the potential biological mechanisms of this reduced risk association warrant further investigation. We also found there to be an inverse association with increasing parity which adds evidence to another previous observation of a higher odds of gastroschisis for first births (McNeese, Selwyn, Duong, Canfield, & Waller, 2015).

Strengths of this study included the use of the rich data available in California and the use of random forest to simultaneously explore such data. The California center of the NBDPS (Reefhuis et al., 2015) utilizes a standardized questionnaire to collect a wide array of data including some paternal factors. These data include many parental exposures and activities experienced periconceptionally. Geocoded residences also allowed integration of previously collected detailed pesticide and air pollution data (Padula et al., 2013; Shaw et al., 2014) to be analyzed with the parental factors. The complete case ascertainment from the well-established California Birth Defects Monitoring Program offered a relatively large sample of cases and enabled stratification by maternal age. Use of random forest, as a data-mining approach, was a strength owing to its non-hypothesis driven variable selection and its capability to examine a large number of variables, accounting for all others simultaneously. Limitations of this study included the case–control nature of the study and the potential for recall bias as well as the moderate participation percentages of case and control mothers. Analyses with a large number of variables are also susceptible to collinearity among the variables and while random forest is designed to account for collinearity, standard regression analyses are not. Lastly, due to the nature of data mining such as random forest, there are concerns regarding multiple testing and spurious associations. It is possible that some of the observed associations are due to chance alone.

Despite the unique epidemiologic profile that characterizes gastroschisis, that is, highly elevated risk for teenage mothers, and an increasing prevalence over the past several decades, the identification of explanatory factors for this profile has been only modestly productive. Here we seemingly explored [idiomatically] *everything but the kitchen sink* and despite our novel and comprehensive approach to generate new hypotheses regarding its unique epidemiology, we did not discover many previously unreported associations for gastroschisis. Our results do add evidence to associations observed with many of these proposed risk factors, particularly given such factors were modeled simultaneously with many of the other proposed risk factors. Our current methods for identifying clues to the unique epidemiology of gastroschisis are perhaps in need of a paradigm shift from the usual suspects of birth defects inquiry that relies on what investigators think is important to ask women to recall to a much deeper biologic inquiry that relies on various biomarkers and myriad omics to characterize pregnancies with and without fetuses with gastroschisis.

ACKNOWLEDGMENTS

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position

of the Centers for Disease Control and Prevention or the California Department of Public Health. We thank the California Department of Public Health, Maternal Child and Adolescent Health Division for providing surveillance data from California for this study. We are grateful for the scientific contributions on pesticide exposure assessment made by Drs. Paul English and Eric Roberts. The authors declare that they have no conflicts of interest.

ORCID

Kari A. Weber https://orcid.org/0000-0001-5138-7389

REFERENCES

Ahrens, K. A., Anderka, M. T., Feldkamp, M. L., Canfield, M. A., Mitchell, A. A., & Werler, M. M. (2013). Antiherpetic medication use and the risk of gastroschisis: Findings from the National Birth Defects Prevention Study, 1997-2007. *Paediatric and Perinatal Epidemiology*, *27*(4), 340–345.

Alwan, S., Reefhuis, J., Rasmussen, S. A., Olney, R. S., & Friedman, J. M. (2007). Use of selective serotonin-reuptake inhibitors in pregnancy and the risk of birth defects. *The New England Journal of Medicine*, *356*(26), 2684–2692.

Canfield, M. A., Honein, M. A., Yuskiv, N., Xing, J., Mai, C. T., Collins, J. S., … Kirby, R. S. (2006). National estimates and race/ethnic-specific variation of selected birth defects in the United States, 1999-2001. *Birth Defects Research Part A, Clinical and Molecular Teratology*, *76*(11), 747–756.

Castilla, E. E., Mastroiacovo, P., & Orioli, I. M. (2008). Gastroschisis: International epidemiology and public health perspectives. *American Journal of Medical Genetics Part C, Seminars in Medical Genetics*, *148c*(3), 162–179.

Croen, L. A., Shaw, G. M., Jensvold, N. G., & Harris, J. A. (1991). Birth defects monitoring in California: A resource for epidemiological research. *Paediatric and Perinatal Epidemiology*, *5*(4), 423–427.

Draper, E. S., Rankin, J., Tonks, A. M., Abrams, K. R., Field, D. J., Clarke, M., & Kurinczuk, J. J. (2008). Recreational drug use: A major risk factor for gastroschisis? *American Journal of Epidemiology*, *167*(4), 485–491.

Feldkamp, M. L., Carmichael, S. L., Shaw, G. M., Panichello, J. D., Moore, C. A., & Botto, L. D. (2011). Maternal nutrition and gastroschisis: Findings from the National Birth Defects Prevention Study. *American Journal of Obstetrics and Gynecology*, *204*(5), 404.e401–404.e410.

Feldkamp, M. L., Krikov, S., Botto, L. D., Shaw, G. M., & Carmichael, S. L. (2014). Better diet quality before pregnancy is associated with reduced risk of gastroschisis in Hispanic women. *The Journal of Nutrition*, *144*(11), 1781–1786.

Feldkamp, M. L., Meyer, R. E., Krikov, S., & Botto, L. D. (2010). Acetaminophen use in pregnancy and risk of birth defects: Findings from the national birth defects prevention study. *Obstetrics and Gynecology*, *115*(1), 109–115.

Gill, S. K., Broussard, C., Devine, O., Green, R. F., Rasmussen, S. A., & Reefhuis, J. (2012). Association between maternal age and birth defects of unknown etiology: United States, 1997-2007. *Birth Defects Research Part A, Clinical and Molecular Teratology*, *94*(12), 1010–1018.

Interrante, J. D., Ailes, E. C., Lind, J. N., Anderka, M., Feldkamp, M. L., Werler, M. M., … Broussard, C. S. (2017). Risk comparison for prenatal use of analgesics and selected birth defects, national birth defects prevention study 1997-2011. *Annals of Epidemiology*, *27*(10), 645–653 e2.

Jones, A. M., Isenburg, J., Salemi, J. L., Arnold, K. E., Mai, C. T., Aggarwal, D., … Honein, M. A. (2016). Increasing prevalence of Gastroschisis--14 states, 1995-2012. *MMWR Morbidity and mortality weekly report*, *65*(2), 23–26.

Khodr, Z. G., Lupo, P. J., Canfield, M. A., Chan, W., Cai, Y., & Mitchell, L. E. (2013). Hispanic ethnicity and acculturation, maternal age and the risk of gastroschisis in the national birth defects prevention study. *Birth Defects Research Part A, Clinical and Molecular Teratology*, *97*(8), 538–545.

Lam, P. K., Torfs, C. P., & Brand, R. J. (1999). A low pregnancy body mass index is a risk factor for an offspring with gastroschisis. *Epidemiology (Cambridge, Mass)*, *10*(6), 717–721.

Lammer, E. J., Iovannisci, D. M., Tom, L., Schultz, K., & Shaw, G. M. (2008). Gastroschisis: A gene-environment model involving the VEGF-NOS3 pathway. *American Journal of Medical Genetics Part C, Seminars in Medical Genetics*, *148c*(3), 213–218.

Lin, S., Munsie, J. P., Herdt-Losavio, M. L., Bell, E., Druschel, C., Romitti, P. A., & Olney, R. (2008). Maternal asthma medication use and the risk of gastroschisis. *American Journal of Epidemiology*, *168*(1), 73–79.

McNeese, M. L., Selwyn, B. J., Duong, H., Canfield, M., & Waller, D. K. (2015). The association between maternal parity and birth defects. *Birth defects research Part A, Clinical and molecular teratology*, *103*(2), 144–156.

Padula, A. M., Tager, I. B., Carmichael, S. L., Hammond, S. K., Lurmann, F., & Shaw, G. M. (2013). The association of ambient air pollution and traffic exposures with selected congenital anomalies in the San Joaquin Valley of California. *American Journal of Epidemiology*, *177*(10), 1074–1085.

Parker, S. E., Mai, C. T., Canfield, M. A., Rickard, R., Wang, Y., Meyer, R. E., … Correa, A. (2010). Updated national birth prevalence estimates for selected birth defects in the United States, 2004-2006. *Birth Defects Research Part A, Clinical and Molecular Teratology*, *88*(12), 1008–1016.

Polen, K. N., Rasmussen, S. A., Riehle-Colarusso, T., & Reefhuis, J. (2013). Association between reported venlafaxine use in early pregnancy and birth defects, national birth defects prevention study, 1997-2007. *Birth Defects Research Part A, Clinical and Molecular Teratology*, *97*(1), 28–35.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rasmussen, S. A., & Frias, J. L. (2008). Non-genetic risk factors for gastroschisis. *American Journal of Medical Genetics Part C, Seminars in Medical Genetics*, *148c*(3), 199–212.

Rasmussen, S. A., Olney, R. S., Holmes, L. B., Lin, A. E., Keppler-Noreuil, K. M., & Moore, C. A. (2003). Guidelines for case classification for the National Birth Defects Prevention Study. *Birth Defects Research Part A, Clinical and Molecular Teratology*, *67*(3), 193–201.

Reefhuis, J., Gilboa, S. M., Anderka, M., Browne, M. L., Feldkamp, M. L., Hobbs, C. A., … Honein, M. A. (2015). The national birth defects prevention study: A review of the methods. *Birth Defects Research Part A, Clinical and Molecular Teratology*, *103*(8), 656–669.

Reefhuis, J., & Honein, M. A. (2004). Maternal age and non-chromosomal birth defects, Atlanta--1968-2000: Teenager or thirty-something, who is at risk? *Birth Defects Research Part A, Clinical and Molecular Teratology*, *70*(9), 572–579.

Richardson, S., Browne, M. L., Rasmussen, S. A., Druschel, C. M., Sun, L., Jabs, E. W., & Romitti, P. A. (2011). Associations between periconceptional alcohol consumption and craniosynostosis, omphalocele, and gastroschisis. *Birth Defects Research Part A, Clinical and Molecular Teratology*, *91*(7), 623–630.

Shaw, G. M., Yang, W., Roberts, E., Kegley, S. E., Padula, A., English, P. B., & Carmichael, S. L. (2014). Early pregnancy agricultural pesticide exposures and risk of gastroschisis among offspring in the San Joaquin Valley of California. *Birth Defects Research Part A, Clinical and Molecular Teratology*, *100*(9), 686–694.

St Louis, A. M., Kim, K., Browne, M. L., Liu, G., Liberman, R. F., Nembhard, W. N., … Kirby, R. S. (2017). Prevalence trends of selected major birth defects: A multi-state population-based retrospective study, United States, 1999 to 2007. *Birth Defects Research Part A, Clinical and Molecular Teratology*, *109*(18), 1442–1450.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348.

Torfs, C. P., Lam, P. K., Schaffer, D. M., & Brand, R. J. (1998). Association between mothers' nutrient intake and their offspring's risk of gastroschisis. *Teratology*, *58*(6), 241–250.

Torfs, C. P., Velie, E. M., Oechsli, F. W., Bateson, T. F., & Curry, C. J. (1994). A population-based study of gastroschisis: Demographic, pregnancy, and lifestyle risk factors. *Teratology*, *50*(1), 44–53.

Vu, L. T., Nobuhara, K. K., Laurent, C., & Shaw, G. M. (2008). Increasing prevalence of gastroschisis: Population-based study in California. *The Journal of Pediatrics*, *152*(6), 807–811.

Waller, D. K., Gallaway, M. S., Taylor, L. G., Ramadhani, T. A., Canfield, M. A., Scheuerle, A., … Correa, A. (2010). Use of oral contraceptives in pregnancy and major structural birth defects in offspring. *Epidemiology (Cambridge, Mass)*, *21*(2), 232–239.

Waller, D. K., Hashmi, S. S., Hoyt, A. T., Duong, H. T., Tinker, S. C., Gallaway, M. S., … Canfield, M. A. (2018). Maternal report of fever from cold or flu during early pregnancy and the risk for noncardiac birth defects, National Birth Defects Prevention Study, 1997-2011. *Birth Defects Research Part A, Clinical and Molecular Teratology*, *110*(4), 342–351.

Waller, D. K., Shaw, G. M., Rasmussen, S. A., Hobbs, C. A., Canfield, M. A., Siega-Riz, A. M., … Correa, A. (2007). Prepregnancy obesity as a risk factor for structural birth defects. *Archives of Pediatrics & Adolescent Medicine*, *161*(8), 745–750.

Werler, M. M., Mitchell, A. A., & Shapiro, S. (1992). Demographic, reproductive, medical, and environmental factors in relation to gastroschisis. *Teratology*, *45*(4), 353–360.

Werler, M. M., Sheehan, J. E., & Mitchell, A. A. (2002). Maternal medication use and risks of gastroschisis and small intestinal atresia. *American Journal of Epidemiology*, *155*(1), 26–31.

Willett, W. C., Sampson, L., Stampfer, M. J., Rosner, B., Bain, C., Witschi, J., … Speizer, F. E. (1985). Reproducibility and validity of a semiquantitative food frequency questionnaire. *American Journal of Epidemiology*, *122*(1), 51–65.

Yazdy, M. M., Mitchell, A. A., & Werler, M. M. (2014). Maternal genitourinary infections and the risk of gastroschisis. *American Journal of Epidemiology*, *180*(5), 518–525.

How to cite this article: Weber KA, Yang W, Carmichael SL, Padula AM, Shaw GM. A machine learning approach to investigate potential risk factors for gastroschisis in California. *Birth Defects Research*. 2019;111:212–221. https://doi.org/10.1002/bdr2.1441

APPENDIX A: Variables from the California Center Included in Random Forest

Variables were included if they were in both versions of the computer-assisted telephone interview, occurred during the time-period from 1 month before pregnancy (B1) until the second month of pregnancy (P1–P2) if time-varying, and had a frequency of ≥0.1%. Variables included both categorical and continuous responses.

Categorical variables included: maternal and paternal education (<high school, high school, >high school), maternal and paternal race/ethnicity (white non-Hispanic, Hispanic foreign-born, Hispanic US-born, black non-Hispanic, other), maternal feelings about pregnancy (wanted to be pregnant, wanted to wait until later, did not care, pregnant despite consistent contraceptive use), infant sex (male, female), annual household income (<$10,000, $10,000–$50,000, >$50,000), timing of pregnancy discovery (first trimester, second/third trimester), and pre-pregnancy obesity (<30 kg/m2, ≥30 kg/ m2).

Continuous variables included: number of people supported with household income, timing of first prenatal visit (month), maternal and paternal age at delivery (years), parity, number of previous pregnancies, miscarriages, caffeine from coffee (mg/day), caffeine from tea (mg/day), caffeine from soda (mg/day), total caffeine (mg/day), time spent per shower (minutes), time spent per bath (minutes), and number of jobs from B1–P2.

Variables with responses of yes or no fell into two categories, overall and by time-period (separate responses for B1, P1, and P2). Overall yes/no variables included: pesticide exposure (Dichlorophenoxy acid or ester, Alcohol/Ether, Alkyl Phthalate, Amide, Aryloxyphenoxy propionic acid, Avermectin, Azole, Bipyridylium, Hydroxybenzonitrile, Insect Growth Regulator, Chloroacetanilide, Chlorinated Phenol, Copper-containing compound, Cyclohexenone derivative, 2,6-Dinitroaniline, Diacylhydrazine, Benzoic acid, Dicarboximide, Dithiocarbamate, Endothall, Phosphonoglycine, Glyco Ether, Halogenated organic, Imidazolinone, Monochlorophenoxy acid or ester, Dithiocarbamate, N-Methyl Carbamate, Neonicotinoid, Organochlorine, Organophosphate, Organoarsenic, Bis-Carbamate, Petroleum derivative, Phenol, Piperonyl, Polyalkyloxy Compound, Pyrethroid, Pyridazinone, Quaternary Ammonium Compound, Silicone, Streptomycin, Strobin, Sulfonylurea, Thiocarbamate, Pyridinecarboxylic acid, Benzimidazole, Thiophthalimide, Triazine, Chloropyridinyl, Urea, Xylylalanine, and Inorganic-Zinc), folic acid-containing vitamin intake, singleton birth, chorionic villus sampling, pre-pregnancy diabetes, use of insulin, use of fertility medication or procedure, medication for pregnancy nausea, pre-pregnancy high blood pressure, high blood pressure during pregnancy, epilepsy, seizures, medication for seizures, beer intake, wine intake, mixed drink intake, shots of liquor intake, other drink intake, maternal and paternal active military duty, any household participation in occupational pesticide application, maternal and paternal health problems or birth defects diagnosed in childhood, other relative health problems or birth defects diagnosed in childhood, private well drinking water source, and father employed.

Each time-period variable had a separate variable for B1, P1, and P2. These variables included CT/CAT scan, MRI, X-ray, other X-ray or scan, urinary tract infection, pelvic inflammatory disease, other illness, surgery, birth control pill use, other birth control use, pregnancy nausea (P1 and P2 only), cereal intake, food supplement intake, cigarette smoking, smoking in the household, smoking in the workplace or school, alcohol use, hot tub/Jacuzzi/sauna use, and maternal and paternal substance abuse. Substance abuse combined use of marijuana, hash, cocaine, crack, hallucinogens, heroin, mushrooms, and other. Use of non-steroidal anti-inflammatory (aspirin, ibuprofen, ibuprofen [lysine salt], naproxen, and naproxen sodium), acetaminophen, Nitrofurantoin (nitrofurantoin, nitrofurantoin monohydrate, and nitrofurantoin sodium), anti-depressant (selective serotonin reuptake inhibitors or bupropion), and benzodiazepine were also included as time-period variables with a response of yes or no.

Nutrient and air pollution values were divided into three categories based on values among controls: those with exposure levels <25%, those within the interquartile range, and those with exposure levels >75%. Nutrient variables

included dietary intake of: fat (g), total carbohydrate (g), total protein (g), alanine (g), methionine (g), cysteine (g), total choline (mg), betaine (mg), calcium (mg), alpha-carotene (μg), beta-carotene (μg), copper (mg), folate (μg, dietary folate equivalents), iron (mg), lutein (μg), magnesium (mg), niacin (mg), retinol (μg), riboflavin (mg), selenium (μg), thiamin (mg), vitamin A (μg, Retinoic Acid Equivalents), vitamin B6 (mg), vitamin B12 (μg), vitamin C (mg), zinc (mg), glycemic index, and diet quality index (quartiles). Air pollution variables included daily 24-hr averages of nitrogen oxide, nitrogen dioxide, particulate matter <10 μm (PM10), particulate matter <2.5 μm (PM2.5), and carbon monoxide, and a daily 8-hr maximum of ozone and averaged over the study period. Traffic density was included as a dimensionless indicator based on traffic volumes within a 300-m radius.

All food frequency questionnaire items were divided into the following categories: <1x per month, 1–3x per month, 1x per week, 2–6x per week, or > 6x per week. Variables from the food frequency questionnaire, and the units of measurement, included: skim or low fat milk (8 oz. glass); whole milk (8 oz. glass); yogurt (1 cup); ice cream (1/2 cup); cottage or ricotta cheese (1/2 cup); other cheese (1 slice or 1 oz. serving); margarine (pat); butter (pat); apples or pears (1); oranges (1); orange juice (1 glass); peaches, apricots, plums, or nectarines (1 fresh or 1/2 cup canned); bananas (1); other fruits, fresh, frozen, or canned (1/2 cup); tomatoes (1) or tomato juice (small glass); string beans (1/2 cup); broccoli (1/2 cup); cabbage, cauliflower, or brussel sprouts (1/2 cup); carrots, raw (1/2 carrot or 2–4 sticks); carrots, cooked (1/2 cup); corn (1 ear or 1/2 cup frozen, canned); peas or lima beans (1/2 cup frozen, canned); yams or sweet potatoes (1/2 cup); spinach or collard greens, cooked (1/2 cup); beans or lentils, baked or dried (1/2 cup); yellow squash (1/2 cup); eggs (1); chicken or turkey (4–6 oz.); bacon (2 slices); hot dogs (1); processed meats, for example, sausage, salami, bologna, chorizo, and so forth (piece or slice); liver (3–4 oz.); hamburger (1 patty); beef, pork, lamb or cabrito as a sandwich or mixed dish, for example, stew, casserole, lasagna, and so forth; beef, pork, lamb or cabrito as a main dish, for example, steak, roast, ham, and so forth (4–6 oz.); fish (3–5 oz.); chocolate (1 oz.); candy without chocolate (1 oz.); pie (slice); cookies (1); white bread (slice), including pita bread; dark bread (slice), including wheat pita bread; french fried potatoes (4 oz.); potatoes, baked, boiled (1) or mashed (1 cup); rice or pasta for example, spanish rice, spaghetti, noodles, and so forth (1 cup); potato chips or corn chips (small bag or 1 oz.); nuts (small packet or 1 oz.); peanut butter (1 tbs); oil and vinegar dressing for example, Italian (1 tbs); cantaloupe (1/4 melon); avocado (1) or guacamole (1 cup); raw chile peppers, jalapeno (1); salsa (1 cup); chicken livers (1 oz.); organ meats, Barbacoa, Menudo, sweetbreads, tongue, intestines (3–4 oz.); tortilla (1); refried beans (1 cup).