January 2013

The Influence Of Maternal Nutrition Intake On The Birth Weight Of Newborns In Lanzhou, China

Weiwei Zhang
Yale University, vera.chang.ww@gmail.com

The Influence of Maternal Nutrition Intake on the Birth Weight of Newborns in Lanzhou, China

By

Weiwei Zhang

A Thesis Presented to

The Faculty of the Department of Biostatistics

Yale School of Public Health

In Candidacy for the Degree of

Master of Public Health

2013

Permission to Copy

Permission for photocopying, microfilming, or computer electronic scanning of "The Influence of Maternal Nutrition Intake on the Birth Weight of Newborns in Lanzhou, China" for the purpose of individual scholarly consultation of reference is hereby granted by the author. This permission is not to be interpreted as affecting publication of this work or otherwise placing it in the public domain, and the author reserves all rights of ownership guaranteed under common law protection of unpublished manuscripts.

_Weiwei Zhang_____
Signature of Author

_April 30, 2013_____
Date

ABSTRACT

Background: The objective of this paper is to evaluate the role of maternal nutrition intake before and during pregnancy on infant birth weight. Self-reported data on intake of 94 nutrients during four time periods (one year before pregnancy and during the three trimesters of pregnancy) was collected from over 7,000 women during 2010-2012 in Gansu Provincial Maternity and Child Care Hospital in Lanzhou, China.

Methods: Hierarchical clustering and k-means clustering were performed to detect any potentially influential nutrients. Principal Component Analysis (PCA) was used to reduce the dimension of predictors and linear regression analysis was subsequently conducted to fit new nutrient components and confounders on birth weight. Since the Globaltest showed significant influence of nutrient intake in participant group with low pre-pregnancy maternal Body Mass Index (BMI), stepwise selection and Least Absolute Shrinkage and Selection Operator (Lasso) methods were then performed directly over the 94 nutrients in low BMI group.

Results: Hierarchical clustering and k-means clustering both resulted in six clusters, but the averages of birth weight in the six clusters were not significantly different from each other based on ANOVA result (p-value=0.1229 and 0.1032, for hierarchical and k-means clustering, respectively). Principal Components Analysis (PCA) was then performed and selected 10 new nutrient components to represent the original 94 nutrients. The following step was to fit the 10 new components and 9 confounders on birth weight in a linear regression model. This procedure was repeated for the four time periods. For each time period, there were two components showing strong association with birth weight.

Comparing to the base model with only confounders, full models with nutrient components had very small R-squared increase (around 0.01). According to the Globaltest result, the overall effect of 94 nutrients showed strong association with birth weight in low pre-pregnancy maternal BMI group in all four time periods (p-values=0.0439, 0.0033, 0.0017, and 0.0024, for four time periods, respectively). In the model selection part, stepwise selection resulted in two significant nutrient variables out of 94 variables for each time period. Specifically, variable 'insoluble dietary fiber' showed strong association with birth weight (p-value=0.0022, 0.0051 and 0.0062, for pre-pregnancy, the 2^{nd} and 3^{rd} trimester, respectively). In addition, a few vitamin nutrients showed strong association with birth weight (p-values < 0.0001), but the estimated coefficients were very small. Lasso method showed similar results as stepwise selection. Few nutrients showed significant influence but with very small estimated coefficients, and R-squared of the full models had very small increase compared to the base model with only confounders.

Conclusions: Several methods have been tried to test the association between maternal nutrition intake and birth weight of newborns, such as clustering on predictors and observations, Globaltest, stepwise selection and Lasso method in linear regression. These methods showed consistent results that overall maternal nutrition intake was significantly associated with infant birth weight in low maternal BMI group, and a few individual nutrients showed significant association. It is suggested that instead of focusing on altered consumption of individual nutrients, overall maternal nutrition intake should be improved to help control birth weight in the normal range.

TABLE OF CONTENTS

INTRODUCTION

Birth weight of newborns is an important indicator of infant health condition as well as their adulthood health development. Extreme value of birth weight, such as small for gestational age or big for gestational age, may lead to higher risk of cardiovascular, diabetes, obesity and other diseases. Many factors may influence birth weight, including gestational age, prenatal growth rate, and other genetic and environmental factors. Studies have shown that maternal nutrition intake is also a very influential factor, especially among malnourished maternal group. It is suggested that improved nutritional intake on certain nutrient or certain type of nutrient group may help reduce the rate of low birth weight and balanced nutrition intake may help control high birth weight rate.

The Influence of Birth Weight Over Health Conditions in Later Life Stages

The body weight of newborns, i.e. birth weight, has important influence over infants' health as well as their later life stages. Several studies have been conducted to evaluate the association between people's birth weight and their childhood, adolescent or adulthood health conditions. Many studies have focused on extreme birth weight, including low birth weight and high birth weight, and its negative influence over health conditions. It has been shown that low or high birth weight can both lead to increased risk of obesity (Barker, et al., 1997; Malina, et al., 1996; Ounsted, et al., 1984) and cardiovascular disease (Rich-Edwards, et al., 1997; Singhal, et al., 2003) in adolescence as well as adulthood. Several studies on adult women have suggested a link between their low birth weight and an increased risk for diabetes (Curhan, et al., 1996; McCance, et al.,

1994; Rich-Edwards, et al., 1999). Birth weight is also found to be associated with breast cancer (Michels, et al., 1996). On the other hand, within normal birth weight range, birth weight could still influence individuals' performance as a teenager or an adult (Matte, et al., 2001; Record, et al., 1969; Richards, et al., 2002). The association between birth weight and intelligence level was assessed among seven-year-old children who were born with normal birth weight in 12 cities of the United States. The study showed that Intelligence Score increased with birth weight in a linear fashion (Matte, et al., 2001). Given the influence of birth weight discussed above, we can see that it affects many aspects of people's health condition. Hence it is of significant importance to discover the risk factors that would affect birth weight.

Potential Risk Factors for Birth Weight

There are many biological, genetic and environmental factors that could influence infants' birth weight (Fuster, et al., 2013; Kramer, 1987; Wells, 2002). Gestational age and prenatal growth rate are acknowledged to be the two most direct and influential determinants of birth weight. Gestation is usually around 40 weeks in humans and includes three trimesters. Studies have shown that longer gestational age usually has higher birth weight (Donahue, et al., 2010; Zhang and Bowes, 1995). Demographic and socioeconomic factors are also contributors to birth weight, though the trend may change in different populations (Kramer, et al., 2000; Mohsin, et al., 2003; Parker, et al., 1994). For instance, higher education is positively associated with birth weight among native Greek women, but negatively associated among the immigrant mothers (Tsimbos and Verropoulou, 2011). In addition, birth weight determinants may vary among countries

with different economic levels. Cigarette smoking is the most influential factor for birth weight in developed countries, followed by nutritional factors and pre-pregnancy weight, while racial origins, gestational nutrition and low pre-pregnancy weight play the most important roles in developing countries (Bonellie, 2012). Many studies have shown that maternal pre-pregnancy weight is closely associated with newborns' birth weight (Campbell, et al., 2012; Hunt, et al., 2013; Susser and Stein, 1982). Furthermore, other than the risk factors discussed above, another important focus of determinants on birth weight is maternal nutrition intake.

Nutrition Intake Influence Over Birth Weight

A lot of interests have been paid to micronutrients. Micronutrients are nutrients required by humans and other organisms throughout life in small quantities, including iron, iodine, manganese, zinc, and many different kinds of vitamins, such as folic acid and niacin. Some studies have shown that multiple micronutrients supplementation is associated with an increase in mean birth weight, as well as reductions in the prevalence of low birth weight and small for gestational age. This association also depends on the amount of micronutrients supplements and the time of initiation. In a community with poor economic conditions in Nepal, multiple micronutrient supplementations during the second and third trimester led to reductions in the prevalence of low birth weight (Osrin, et al., 2005). Multiple micronutrients supplementation performs better in reducing the risk of low birth weight than iron-folic acid supplementation alone (Zagre, et al., 2007). In addition, relatively longer intake period resulted in better birth weight outcome.

However, another study conducted in Pakistan concluded that multiple micronutrients have only modest effect on birth weight (Bhutta, et al., 2009).

At the individual level for micronutrients, calcium supplementation shows protection against low birth weight (Merialdi, et al., 2003), especially for women at a high risk of gestational hypertension and were at malnutrition in calcium before pregnancy (Atallah, et al., 2002). However, studies examining other single nutrient effect did not find much significant connection to birth weight. For instance, no convincing evidence showed that zinc supplementation resulted in better infant outcome (Mahomed, et al., 2007; Mori, et al., 2012; Tamura, et al., 1997). Similarly, no detectable effect of iron was found on birth weight (Mahomed, 2000). In addition, a prospective cohort study was conducted in Ethiopia to investigate the interact influence of prenatal zinc and vitamin A over birth weight, but no association was found (Gebremedhin, et al., 2012).

Many studies have been focused on the vitamin group specifically. It has been shown that the regular periconceptional multivitamin use may reduce the risk of low birth weight, but only in women with a pre-pregnancy BMI in the normal range (Catov, et al., 2011). Vitamin A influence has been examined in several studies but with contradictable conclusions. Vitamin A supplementation was found not to affect the risk of low birth weight in a study of 289 pregnant women in Birmingham, Alabama, USA (Tamura, et al., 1997). In contrast, in a study conducted in Israel, low vitamin A intake was found to be more frequent in low birth weight infants (Gorodischer, et al., 1995). The different outcomes in the association between vitamin A and birth weight may be due to the

regional difference. Several studies have tested the effect of maternal vitamin D status over birth weight, resulting in conflicting conclusions as well. Some observational studies with small sample size (N<600) did not show any relation between maternal vitamin D intake and birth weight (Akcakus, et al., 2006; Gale, et al., 2008; Morley, et al., 2009). However, some large studies found that vitamin D intake could reduce the risk of low birth weight (Bodnar, et al., 2010; Prentice, et al., 2009). No association was found between maternal vitamin E intake and birth weight (Merialdi, et al., 2003; Tamura, et al., 1997). The effect of vitamin C intake during pregnancy over birth weight was not significant either (Dror and Allen, 2012; Rumbold and Crowther, 2005).

In addition to the common vitamins discussed above, a type of vitamin B, folic acid, has been paid a lot of attention to. Folic acid concentrations showed positive relationships with the birth weight of infants (Tamura, et al., 1997). Furthermore, a combined effect of iron and folic has been examined. Iron-folic acid has shown to be associated with an increase in birth weight in less developed countries and this effect was greater among women with higher BMI (Fall, et al., 2009). A study conducted in rural areas of China also indicated that iron-folic acid supplementation has larger effect in birth weight than folic acid alone (Zeng, et al., 2008).

Energy and protein do not seem to significantly influence birth weight (Kramer and Kakuma, 2003). Diet high in carbohydrates and fat does not lead to increase in birth weight (Bouwland-Both, et al., 2013). However, some other studies have pointed out that the increase in birth weight contributed by calories and protein depends on the existing

dietary pattern of the population (Lechtig, et al., 1975). It has also indicated that nutrition intake has bigger influence among mothers with poor pre-pregnancy nutritional status than other pregnant women. Insufficient consumption of nutrition may result in low birth weight. In contrast, balanced protein and energy supplementation has proved to be able to reduce the risk of low birth weight (Merialdi, et al., 2003). However, on the other hand, the negative effect of protein intake on birth weight was also detected in a trial conducted in New York. At the same time, some other studies have also shown no significant correlation with birth weight (Tamura, et al., 1997). A study on 847 Dutch women has shown that no association was revealed between high carbohydrates and fat intake and birth weight (Timmermans, et al., 2012).

According to the discussion above, it has presented that the results of a lot of nutrition intake studies are not consistent. This inconsistency may be due to regional difference, population difference, and various sample sizes and study designs. Most of the studies usually choose a few nutrients to focus on instead of the overall nutrition influence. Though some review papers have examined more types of nutrients, their conclusions are mostly based on pooling results from different trials together, which may lead to bias and imprecision of the conclusions. Furthermore, most studies were conducted among western populations. Our study has covered almost all kinds of nutrients (94 nutrients), and the time period has covered from one year before pregnancy until the completion of three pregnancy trimesters. The study mainly focused on Chinese population with the sample size being over 7,000 participants. Data was originally recorded as the food intake amount, and then the information was converted into nutrition intake amount by

combining food nutritional information table. Multiple analytical technics including clustering, globaltest and linear regression modeling have been utilized to detect the individual and overall effects of maternal nutrition intake over birth weight.

METHODS

The birth cohort study was carried out in Gansu Provincial Maternity and Child Care Hospital in Lanzhou, Gansu, China (See Figure 1 for the geographic information of Lanzhou, China). Information on nutrition intake one year before pregnancy and during three trimesters of pregnancy was collected from over 7,000 pregnant women from year 2010 to 2012 through a food frequency questionnaire. Information on parental demographic, maternal health conditions before and during pregnancy, parental life style factors, maternal diet habits and so on was collected through in-person interview using a standardized structured questionnaire. In our study, we focused on food intake information of 33 different categories of foods, adjusting for maternal pre-pregnancy health condition and parental demographic information as confounders.

The objective of this paper is to conduct an evaluation on the role of nutritional intake before and during pregnancy on infant birth weight. We are interested in seeing if certain nutrient or certain type of nutrients group would have significant influence on birth weight of newborns. The analysis in our study was limited to singleton live births without birth defects. These exclusions left us with 7376 observations. Our response variable is birth weight measured in grams (continuous variable). Our predictors of interest are 94 nutrient variables, including vitamins, proteins, minerals, fats and water. Given what we

have discussed in the introduction section, maternal health conditions and parental demographic information have significant influence over infants' birth weight. Hence gestational age, maternal age, pre-pregnancy BMI, education level, family income, nationality, smoking status, hypertension status and diabetes status were adjusted as confounders in the analysis.

As mentioned in previous section, maternal pre-pregnancy weight is closely associated with newborns birth weight. Body mass index (BMI) is an important indicator of body fatness for most people and provides weight categories related to health problems (Center for Disease Control and Prevention, USA, 2011). BMI (kg/m^2) is calculated from weight (kg) divided by the square of height (m^2). It is usually acknowledged that the upper bound of normal BMI is 25 kg/m^2. However, World Health Organization (WHO) has modified the threshold of being overweight for Asian population of BMI from 25 kg/m^2 to 23 kg/m^2 based on some studies conducted on Asian women (Choo, 2002). The total cohort was divided into three subgroups based on maternal pre-pregnancy BMI level, that is, low BMI group (BMI<18.5 kg/m^2), normal BMI group (18.5 $kg/m^2 \leq$ BMI $\leq$ 23 kg/m^2) and High BMI group (BMI > 23 kg/m^2).

Our analysis strategies have covered various technics, and we examined maternal nutrition intake effect in the total cohort as well as among three subgroups based on maternal pre-pregnancy BMI level. Descriptive analysis was first performed to obtain the general information of maternal demographics and health conditions. These variables were adjusted as confounders in our study. Hierarchical clustering and k-means clustering

were implemented over observations based on nutrients to detect any potential influential nutrients. Then Globaltest was conducted to test the overall influence of nutrients in subgroups divided by pre-pregnancy maternal BMI level and pregnancy time periods. Marginal analysis of each nutrient was also performed to detect individual nutrient effect. Linear regression model selection by stepwise and Lasso method were then conducted to detect significant nutrient variables and compare models.

Descriptive Analysis

In our data set, the response variable birth weight has mean value of 3282.26 grams, ranging from 600 grams to 5500 grams (N=7366) and is approximately normal distributed (skewness=-0.78). The average maternal age is 28.99 years (N=7374, STD=4.35) and normally distributed (skewness=0.44, range=14.6-49.4 years). The mean pregnancy body mass index (BMI) is 20.57 kg/m^2 (N=7153, STD=2.63) and slightly right skewed (skewness=0.99, range=13.34-38.10 kg/m^2). The gestational age has an average value of 38.59 weeks (N=7372, STD=1.83) and shows some evidence of left skewed (skewness=-2.25, range=26-44 weeks). Education level ranges from less than middle school level to college level, and almost 60% of the participants have education level equal to or higher than college. Family income ranges from less than 2,000 CNY per month to more than 4,000 CNY per month, with almost 50% of families having income between 2,000 and 4,000 CNY. 91.73% of the participants have nationality of Han. With regards to maternal health conditions, 20.62% of the participants have either smoked or been exposed to smoking environment during pregnancy, 4.43% of them experienced either maternal preeclampsia or gestational hypertension and 0.66% have diabetes.

Descriptive analysis in the total cohort and three subgroups of maternal demographics information and health condition were given in Table 1, and histograms of continuous variables could be referred to in Figure 2.

Data Merging

The original data was recorded in two separate data sets:

(1) The first data set contains 7376 participants' food intake of 33 different categories of food during 4 time periods (before pregnancy and three trimesters of pregnancy). Each food has 20 columns containing information of food intake amount on a daily, weekly and monthly basis in the four time periods. Since our research interest is to evaluate the association between maternal nutrition intake on a daily basis in different pregnant periods and infants' birth weight, all information was transformed to a daily basis. Therefore, weekly food intake amount was divided by 7 and monthly food intake amount was divided by 30, assuming that there are 30 days in a month. The plots of some food intake amount against birth weight could be referred to in Figure 3.

(2) This data set contains 94 nutrients information for 33 kinds of food based on 100 grams of each food. Since not every food has all the nutrients listed, there are missing data. We coded the missing data as zero. Moreover, some food may contain very small amount of certain nutrients, and we coded the amount as 10 times smaller than the smallest nutrient amount among other nutrients. The nutrition information of different types of food can be found in Table 3.

Matrix multiplication was used to combine the two data sets, thus transforming data as food intake amount to nutrient intake amount. A new data set of 7376 participants' daily nutrition intake on 94 nutrients was obtained. During the transformation process, some estimation of food density and size were adopted. Most food in the first data set was recorded as weight (grams), therefore we calculated how many 100 grams they contained and then multiplied directly with nutrition information table. Though in each food category there may be different types of food with various nutrients amount, such as different kinds of green vegetables may differ in nutrition intake, we used the most common and popular one in Chinese people's diet to estimate the average value. In addition, for some food, more conversion was conducted between food amount and nutrition amount. For instance, egg was recorded as the number of eggs instead of the weight. Therefore we used the estimated average egg weight of 50 grams to calculate the nutrient amount. Milk was recorded in volume, so we use the approximated density of $1kg/m^3$ to estimate the weight of milk.

Data Mining

The pre-knowledge of the data set was very limited. Hence clustering on observations was initially carried out to get a sense of the data set and explore some potential relations between nutrients and birth weight.

The basic idea of cluster analysis is to find groups of observations that are similar to each other within clusters based on the measured variables and dissimilar from other clusters on the same variables (Milligan and Cooper, 1987). There are several clustering methods.

Two most popular ones for clustering over observations are hierarchical clustering and k-means clustering. In our study, maternal nutrition intake (94 nutrients) was the measured characteristics. If nutrients have a significant influence on birth weight, the averages of birth weight in different clusters are expected to be different from each other. Both hierarchical clustering and k-means clustering were carried out in this study.

Hierarchical Clustering Over Observations

Hierarchical clustering is a method of building a hierarchy of clusters (Ward, 1963). The first step in hierarchical clustering is usually to scale variables to make sure that all variables will have the same weight in the clustering process. In our data set, different nutrient variables have quite different variances. For instance, energy intake (Kcal) has mean 1476.41 Kcal with standard deviation 1533.05 Kcal, while Vitamin B6 has mean 0.69 mg and standard deviation 0.68 mg. Therefore scaling is very necessary in order to obtain a more precise clustering outcome.

Second step in hierarchical clustering is to calculate the distances between objects in pairs. Several methods are available to calculate distances based on the type of data, such as Euclidean distance, Manhattan distance and maximum distance (Szekely and Rizzo, 2005). Usually several combinations of these methods will be carried out to decide which distance method gives the most satisfactory results.

The following step is to choose the agglomeration method that will cluster the two nearest points or clusters. Some common grouping methods are complete linkage, single

linkage, average linkage, centroid linkage and Ward's method (Almeida, et al., 2007; Johnson, 1967). Then Tree Plot based on the outcome of clustering is usually generated to help choose the best agglomeration method and the number of clusters. In addition, several other criteria also play an important role in determining the agglomeration method and the number of clusters (Milligan and Cooper, 1985; Symons, 1981). R-squared (RS) explains the total sum of squares over all variables, so RS near 1 is better. Semi-Partial R-Squared (SPR) measures the relative change in within-clusters, thus small SPR means less loss of homogeneity. Small Root-mean-square standard deviation (RMSSTD) means that it is relatively homogenous within the clusters. Cluster Distance (CD) is small for better clustering. According to the Cubic Clustering Criterion (CCC), usually CCC larger than 2 or 3 indicates good clusters (Sarle, 1983).

In our study, the clustering of observations is based on nutrients, so we are expecting to see if different clusters of observations would have significant nutrition intake, and thus find the most influential nutrients. The average nutrition intakes of the four time periods for each participant were used as the nutrient variables. All the nutrient variables were standardized before clustering. *SAS® PROC CLUSTER* was used to perform hierarchical clustering. Since most of the variables in our data set are continuous variables, the method of Euclidean distance was chosen for measuring the distance between points. Several agglomeration methods were tried, and eventually Ward's method was chosen for grouping points or clusters based on clustering results. Six clusters were chosen based on statistic results and Tree Plot (See Table 3). The total sum of squares over all variables explained by the six clusters was 0.6475. However, the averages of birth weight in the six

clusters were not significantly different from each other based on ANOVA test (p-value=0.1229). Hierarchical clustering based on nutrients did not result in clusters with distinguished birth weight values.

K-means Clustering Over Observations

K-means clustering divides all observations into k clusters by minimizing the within clusters sum of square (Bock, 1985; Hartigan, 1985; Pollard, 1981). As a result, each observation will belong to one and only one cluster. K-means clustering is usually preferred for large data set (sample size N>100). One important aspect of k-means clustering is to select a set of k points as the initial means of clusters. Each observation is assigned to the closest means to form temporary clusters. Then new means of clusters will replace the old means. The process is repeated until no changes occur in the clusters (Hartigan, 1985). Since k-means clustering looks for a local optimum, each time it may result in different answers. The final result depends on the initial selected k means.

Standardization of variables before distance calculation is adopted to avoid imbalanced weights of variables. Euclidean method is usually used to measure distance between points or cluster means. Elbow method is generally used to decide the number of clusters. It is also helpful to just run a random number of k values to see which k values provides better results.

The average nutrition intakes during the four time periods for each participant were used in *SAS® PROC FASTCLUS* to perform k-means clustering. Based on Elbow method and

sum of variance criteria, the number of clusters was determined to be six. However, two of the six clusters were dominating, each consisting of over 3,000 observations. Then ANOVA test was performed to test the difference of average birth weight among clusters. The means of birth weight in different clusters were not statistically significantly different (p-value=0.1032). K-means clustering did not provide distinguished groups based on nutrients.

Both hierarchical clustering and k-means clustering did not provide clusters with significantly different average birth weight based on nutrients. As a result, clustering may fail to detect the association between nutrients intake and birth weight or there may not be strong association between these two factors. Therefore we used other methods to further explore the relationship.

Global Test

The globaltest method was originally targeted for testing a group of genes in clinical studies. It could be used for predicting a response variable (noted as Y) from a set of genes (noted as X). The basic generalized linear model is adopted to explain the concept:

$$E(Y_i|\beta) = h^{-1}(\alpha + \sum_{j=1}^{m} x_{ij}\beta_j)$$

where β_j (j=1,…, m) are the coefficients for regression model (Goeman, 2003). The null hypothesis is usually testing if all $\beta_j = 0$. Assuming $\beta_j s$ are a sample from a distribution with zero as the mean and τ^2 as the variance, so τ^2 represents how far the regression coefficients are from zero. Then $\tau^2 I_m$ is the covariance matrix for β_j and when I_m is a $m \times m$ identical matrix all variables are treated equally. Consequently, the null hypothesis

is as equal to testing if τ^2 equals to zero. Score test is usually performed to test the null hypothesis. Globaltest chooses the model based on the distribution of response variable. For instance, in our case, the response variable (birth weight) is a continuous variable with normal distribution, thus the selected model is linear regression.

More generally applied, the globaltest examines the relationship between a group of variables and a response variable. The method is good for adjusting for the effects of potential confounders and multiple comparisons since variables are grouped together. *R*® *globaltest* Package is used to perform this procedure with response variable (birth weight) Y as a vector and predictors (94 nutrients) X as a matrix. For each of the three BMI groups, the global test of the association between 94 nutrients and birth weight were performed for each time period. In the low BMI group, the global effects of maternal nutrition intake in each time period were statistically significantly associated with birth weight (p-value=0.0439, 0.0033, 0.0017, and 0.0024, respectively). In contrast, nutrition intake did not seem to affect much of birth weight among normal or high BMI groups. Hence main focus would be paid to low BMI group in the further analysis.

Marginal Analysis

Globaltest has indicated overall effect of maternal nutrition intake over birth weight in low pre-pregnancy BMI group. In order to further explore the influence of each nutrient individually, marginal analysis was performed in low maternal BMI group. The base model was birth weight fitted by 9 confounders. For each nutrient, there were four variables measuring nutrient intake during four time periods. Hence the full model for

each nutrient was 9 confounders plus 4 nutrient variables fitting on birth weight in a linear regression model. Then perform F-test to compare the base model and full models to test if any nutrient is very influential.

As a result, 29 out of 94 nutrient variables came out with a p-value less than 0.05 in the F-test. However, since we were repeating the F-test for 94 nutrients, the multiplicity may be an issue. Bonferroni correction is a common method to adjust for multiple comparisons. For the number of n tests, in order to maintain a family-wise significance level at α, the adjusted significance level for each individual test should be $\frac{\alpha}{n}$. Hence in our case, the adjusted statistical significance level for F-test became $\frac{0.05}{94} = 0.00532$. Four nutrient variables were significant at a significance level of 0.00532 after using Bonferroni correction. They were the water (g, p-value=0.0029) intake from food (not including drinking water), folate (ug, p-value=0.0023), vitamin C (mg, p-value=0.0028), and potassium (mg, p-value=0.0016).

Then the four full models were further examined individually. Vitamin C intake variable was significant in the 1st trimester (estimated coefficient = 1.59, p-value=0.0057) and 2nd trimester (estimated coefficient = -3.08, p-value=0.0006). It seemed that vitamin C intake has different influences depending on the time periods. During the 1st trimester, an increase of 1 mg in vitamin C intake would lead to an increase of 1.59 g in birth weight. In contrast, during the 2nd trimester, the same increase in vitamin C would result in a decrease of 3.08 g in birth weight. Potassium intake variable was significant in the 1st

trimester and 2nd trimester as well with estimated coefficients of 0.408 (p-value=0.0010) and -0.554 (p-value=0.0017), respectively. Water intake from food variable seemed to have similar pattern, with 1st trimester estimated coefficient of 0.993 (p-value=0.0005), 2nd trimester estimated coefficient of -1.206 (p-value=0.0033). Folate intake resulted in an estimated coefficient of -1.014 (p-value=0.0149) in the 2nd trimester.

Marginal analysis showed the association between certain nutrients and birth weight. A few nutrients indicated significant influence over birth weight. However, the effects were not consistent in different time periods, and the estimated coefficients were very small. Since the globaltest proved overall significant influence of maternal nutrients intake in low pre-pregnancy maternal BMI group, while marginal analysis resulted in a few influential individual nutrients, we may then consider to select a group of nutrients to represent the original 94 nutrients and then test their association with birth weight.

Clustering Over Nutrient Variables

There are 94 nutrient variables in total in the data set. It is unrealistic and imprecise to include all of them in a linear regression model. In addition, there are some redundancies among those variables. Since many of the nutrients were measured under the same structure, some of the variables are correlated with one another. For instance, different types of fat in the fat group are highly correlated with each other; also the mineral group contains high multicollinearity. Therefore, a method to reducing the dimension of predictors is of necessity.

Principal Components Analysis (PCA) is a common variable-reduction procedure based on the variable clustering. PCA will select principal components that could account for most of the variance in the observed variables. Each principal component is a linear combination of optimally weighted observed variables, and new components are uncorrelated with each other. For example, let vector $\mathbf{X}=(X_1, \ldots, X_p)$ denotes variables X_1, …, X_p, then new variables $\mathbf{X'}=(X_1', \ldots, X_P')$ are linear combinations of $\mathbf{X}$, such as:

$$X_1' = \mathbf{w_1 X} = w_{11} X_1 + \ldots + w_{1p} X_P$$

$$X_2' = \mathbf{w_2 X} = w_{21} X_1 + \ldots + w_{2p} X_P$$

$$\vdots$$

$$X_P' = \mathbf{w_P X} = w_{p1} X_1 + \ldots + w_{pp} X_P$$

where $\mathbf{W}=[w_{ij}]$ is the weight of the jth variable for ith principal component, and they are constructed to make $Var(X_1') \geq Var(X_2') \geq \ldots \geq Var(X_P')$, $\boldsymbol{w_i' w_i} = w_{11}^2 + \ldots + w_{1p}^2 = 1$, and $\boldsymbol{w_i' w_j} = w_{i1} w_{1j} + \ldots + w_{ip} w_{pj} = 0$ for $i \neq j$.

Observed variables with large magnitude of weight contribute the most to a particular principal component. Usually weight magnitude larger than 0.5 is considered as a threshold. The number of new components is equal to the number of observed variables. An eigenvalue represents the amount of variance that is accounted for by a given component. Usually principal components are ranked by their eigenvalues, and only the first few components account for meaningful amounts of variance (Hotelling, 1933).

In our study, SAS® *PROC PRINCOMP* was used to perform PCA over 94 nutrient variables in low BMI group to see if any summary variables on nutrition could be found.

Based on the criteria of Eigenvalue larger than 1 and total variance of observed variables explained (Guttman, 1954; Kaiser and Caffrey, 1965), for each time period respectively, 10 principal components were selected and over 90% of the total variance could be explained by the 10 principal components. However, none of the principal components had weight magnitude of observed variables larger than 0.5. This may result from the relatively large number of nutrients and the multicollinearity among the nutrients. It is hard to explain the true meaning of each principal component.

The base model in our study is birth weight (response variable) fitted by 9 confounders. Then for each time period, there is a full model of birth weight fitted by 9 confounders plus 10 selected principal components representing the original 94 nutrient variables. A F-test was followed to compare the full model and the base model. This process was repeated for each time period.

The F-test results for four time periods showed that full models and the base model were statistically significant different (p-value=0.0111, 0.0243, 0.0210, 0.0264 for four time periods, respectively). However, when we inspected each principal component, none of the principal components had estimated coefficient value larger than one, and only two or three principal components were significant at significance level of 0.05 in each full model. In addition, R-squares for full models only had a slight increase of approximately 0.01 compared to the base model. In a further developed model, interactions of principal components with significant p-values were added to the full models. As a result, only in the full model for the 2nd trimester there was one interaction coming out with a significant

p-value (Prin1*Prin10: 1.83, p-value=0.0498). All the other interactions were not significant at significance level of 0.1.

According to the results from principal components analysis, though PCA successfully reduced variable dimension by selecting 10 new nutrient variables to represent the original 94 nutrients, the 10 nutrient components did not show significant influence in the linear regression model fitting on birth weight adjusted for 9 confounders. There may be no association between maternal nutrition intake and birth weight or important nutrient information was lost during the process of PCA. In the next step, we tried to build linear regression model based on the 94 nutrients directly in order to confirm the results.

Stepwise Model Selection

Stepwise regression is a combination of forward selection and backward elimination. At each step, variables will be tested to stay or leave. It is one of the most common and traditional methods to building linear regression model. Our base model is to fit 9 confounders on birth weight. Then for the full models, stepwise selection method was used to select important variables from the 94 nutrient variables given that the 9 confounders are already in the model. Stepwise selection was repeated for four time periods for low BMI group. As a result, during each time period, only two nutrients variables were selected in the model at a significance level of 0.05. Insoluble dietary fiber intake during one-year before pregnancy and the 2^{nd} and 3^{rd} trimester of pregnancy showed significant association with birth weight (p-value = 0.0022, 0.0051 and 0.0062, respectively) with estimated coefficients all around 7. An increase of 1 g of insoluble

dietary fiber would increase birth weight of 7 g. A few vitamins, including thiamin, folate and vitamin C, also showed significant association with birth weight, but their estimated coefficients were very small. Large increase in the amount of vitamins intake might slightly improve birth weight, which may be explained by the fact that vitamins are considered to be micronutrients for human beings. In addition, the increases in R-squared of full models were only approximately 0.01. In summary, stepwise selection successfully selected a few significantly influential nutrients, but some of them showed very small estimated coefficients.

Least Absolute Shrinkage and Selection Operator (Lasso)

Though widely used, stepwise selection method is not always considered to be reliable as an automated method. The ordinary least square estimates often have low bias at the sacrifice of relatively large variance of the predicted value. In addition, p-values may be too low due to multiple comparisons, or R-squared may be biasedly high (Efron, et al., 2004). Hence the method of Least Absolute Shrinkage and Selection Operator (Lasso) for model selection was used to improve the overall accuracy by lowering the variance though increasing bias a little bit.

Lasso selection constrains the sum of the absolute values of the regression coefficients to be smaller than a specified parameter in ordinary least squares regression (Tibshirani, 1996). The parameter estimates are selected to minimize:

$$\arg\min \left\{ \sum_i^N \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to } \sum_j |\beta_j| \leq t,$$

which is equivalent to minimizing

$$\frac{1}{2n}\sum(y_i - \alpha - \beta' x_i)^2 + \lambda \sum |\beta_j|$$

The latter formula is the usual least squares plus a penalty associated with a parameter λ, thus when $\lambda = 0$ it gets back to the ordinal least squares.

In a geometry sense, this shrinkage criterion $\sum(y_i - \beta' x_i)^2$ also equals to residual sum of squares function:

$$(\beta - \hat{\beta})' X' X (\beta - \hat{\beta}) = k$$

which centers at the ordinal least squares estimates as the elliptical contour.

Variables should be standardized to have unit variance and zero mean. One important aspect of Lasso selection is to choose the parameter λ. In the R package *glmnet*, usually a large number of λ were used for estimating coefficients, and then the optimal λ could be selected through cross validation procedure. Some coefficients of variables may be zero when the parameter is small enough, thus those variables with nonzero coefficients are selected (Osborne, et al., 2000; Tibshirani, 1996). Each Lasso parameter will result in different subset of variables.

Lasso method showed very similar results as stepwise selection. Insoluble dietary fiber and a few vitamins showed significant association with birth weight. 1 g increase in insoluble dietary fiber during one-year before pregnancy and the 2nd and 3rd trimesters would increase birth weight of 7~8 g (p-value = 0.0004, 0.0021 and 0.0030, respectively). Increasing folate intake of 1 ug would result an increase in birth weight of

around 0.4 g during the 2nd and 3rd trimesters (p-value = 0.001 and 0.0036, respectively). This small influential amount of folate may be due to that folate is considered to be micronutrients for human beings.

RESULTS

Neither hierarchical clustering nor k-means clustering resulted in distinct clusters of birth weight based on nutrients (p-value=0.1229 and 0.1032, for testing birth weight means in clusters from hierarchical and k-means clustering, respectively). Principal Components Analysis (PCA) successfully selected 10 new nutrient components to represent the original 94 nutrients. However, full models of the 10 new components and 9 confounders on birth weight did not show big improvement from base model with only confounders (R-squared increased around 0.01). The Globaltest showed strong overall association of 94 nutrients altogether with birth weight in low pre-pregnancy maternal BMI group in all four time periods (p-values=0.0439, 0.0033, 0.0017, and 0.0024, for four time periods, respectively). No significant association was observed in normal and high pre-pregnancy maternal BMI groups. Stepwise selection resulted in two significant nutrient variables for each time period. Insoluble dietary fiber showed positive association with birth weight (p-value=0.0022, 0.0051 and 0.0062, for pre-pregnancy, the 2nd and 3rd trimester, respectively). In addition, a few vitamin nutrients showed negative association with birth weight (p-values < 0.0001), but the estimated coefficients were very small. Lasso method further confirmed this outcome. Few nutrients (insoluble dietary fiber and several vitamins) showed significant influence, but R-squared of the full models had very small increase compared to base model with only confounders.

DISCUSSION

Advantages of our study

Our study has examined the association between maternal nutrition intake and infant birth weight covering all the three trimesters plus one year before conception. Maternal demographic information and pre-pregnancy health conditions were recorded in details, thus helping adjusted for confounders. Instead of focusing on certain type of nutrients, almost all kinds of nutrients were included in our study (94 nutrients).

In addition to exploring the association in the whole cohort, we tested this association separately in subgroups based on maternal BMI level in order to further detect the subgroup that will benefit the most from improved nutrition intake. It turned out that nutrition intake had the most effect in low maternal BMI group, which may suggest that malnutrition population need to improve overall nutrition intake to elevate infant health condition.

Drawbacks of our study

Since the study was carried out in a single hospital in Gansu, China, the sample may not be representative to general populations. Though measurements for potentially important confounders and nutrients were satisfactory, the sample size was insufficient.

In addition, all the data was collected from self-report questionnaire, bias and imprecision may be of concern. Subjects may not be able to recall the type of food or underestimate

or exaggerate the amount of food intake. This is a common problem in many nutrition studies. The food amount and types are not usually precisely recorded, which may lead to biased study results.

Moreover, when we transformed the intake of food amount to nutrient amount, bias may also occur. For example, we used nutrition information of one type of apple to represent all kinds of apples though their nutrition information may vary from each other. Eggs were recorded in the number of eggs eaten, thus the average weight of an egg was used to estimate the weight of eggs eaten. In addition, when transforming monthly food intake amount into a daily intake amount, we assumed that there are 30 days in a month. This assumption may also lead to imprecision in our study results.

Furthermore, there may be some other methods that should be considered. We mainly focused on linear relations between birth weight and each individual nutrition intake, but quadratic terms of each nutrient or interactions of different nutrients were not taken into consideration. In addition, potential non-linear association between birth weight and nutrients was not analyzed. Other technical methods, such as Independent Component Analysis (ICA), have not been tried due to the limitation of time.

Conclusions

Several methods have been tried to test the association between maternal nutrition intake and birth weight of newborns, such as clustering, global test, stepwise selection and lasso method in linear regression, though limitations existed. Our study indicated maternal

intake of a few nutrients, such as insoluble dietary fiber, folate and vitamin C, are significantly associated with newborns birth weight, but the influence also depends on which time periods of the pregnancy. Insoluble dietary fiber intake during one year before pregnancy and the 2nd and 3rd trimester showed positive association with birth weight. As to vitamins, folate intake during pregnancy showed negative association with birth weight, thiamin intake during one year before pregnancy was negatively associated with birth weight, and vitamin C during the 1st trimester was positively associated with birth weight. The global influence of overall maternal nutrition intake is significant to birth weight among mothers with low pre-pregnancy body mass index. This indicated that nutrition intake seemed to play a more important role among malnutrition mothers. It is suggested that instead of focusing on altered consumption of individual nutrients, overall maternal nutritional intake should be improved to help control birth weight in the healthy range, especially for malnutrition population.

APPENDIX

Tables

Table 1 Descriptive Analysis of Maternal Demographics and Health Conditions

	Total Cohort (N=7376)	LBMI (N=1554)	NBMI (N=4433)	HBMI (N=1166)	P-values
Maternal Age (Years)	28.99±4.35	27.65±3.80	29.08±4.25	30.45±4.60	***<0.0001***
Maternal BMI	20.57±2.63	17.52±0.83	20.47±1.21	24.99±2.01	***<0.0001***
Gestational Age (Weeks)	38.59±1.83	38.63±1.85	38.63±1.77	38.44±1.94	***0.0037***
Maternal Second-hand Smoking (%)	1521 (20.62)	335 (21.78)	893 (20.31)	259 (22.50)	0.1830
Nationality Han (%)	6766 (91.73)	1423 (91.57)	4095 (92.38)	1073 (92.02)	0.5892
Diabetes (%)	49 (0.66)	4 (0.26)	20 (0.45)	24 (2.06)	***<0.0001***
Hypertension (%)	327 (4.43)	38 (2.45)	170 (3.83)	100 (8.58)	***<0.0001***
Education level (%)					***<0.0001***
Middle School or Less	1603 (21.73)	294 (19.10)	917 (20.97)	301 (26.24)	
High School	1282 (17.38)	277 (18.00)	784 (17.93)	196 (17.09)	
Community College	1617 (21.92)	380 (24.69)	957 (21.89)	255 (22.23)	
College	2230 (30.23)	494 (32.10)	1370 (31.34)	331 (28.86)	
Graduate School	508 (6.89)	94 (6.11)	344 (7.87)	64 (5.58)	
Income level (%) (CNY/month)					0.1087
Less than 2,000	1901 (25.77)	391 (26.98)	1096 (26.84)	328 (29.98)	
2,000-3,000	2051 (27.81)	431 (29.74)	1242 (30.41)	343 (31.35)	
3,000-4,000	1428 (19.36)	302 (20.84)	882 (21.60)	226 (20.66)	
More than 4,000	1406 (19.06)	325 (22.43)	864 (21.16)	197 (18.01)	

LBMI: Low BMI group (Pre-pregnancy maternal BMI < 18.5 kg/m^2); NBMI: Normal BMI group (Pre-pregnancy maternal 18.5 $kg/m^2 \leq$ BMI $\leq$ 23 kg/m^2); HBMI: High BMI group (Pre-pregnancy maternal BMI > 23 kg/m^2); p-values: t-test for continuous variables or chi-square test for categorical variables.

Table 2 Partial Nutrition Information for Different Food

Items	Energy (Kcal)	CHO (g)	Insoluble fiber (g)	Thiamin (mg)	Folate (ug)	Vit C (mg)	Fe (mg)	Zn (mg)	Protein (g)
Rice	337.00	78.10	0.00	0.06	11.50	0.00	0.20	1.76	6.40
Flour	354.00	70.90	0.00	0.46	23.30	0.00	6.00	0.20	15.70
Coarse	320.80	78.24	0.00	0.10	16.00	0.00	0.60	1.20	8.24
Pork	395.00	2.40	0.00	0.22	0.00	0.00	1.60	2.06	13.20
Beef	190.00	0.00	0.00	0.03	0.00	0.00	3.20	3.67	18.10
Lamp	198.00	0.00	0.00	0.05	0.00	0.00	2.30	3.22	19.00
Chicken	203.50	0.75	0.00	0.07	0.00	0.00	1.80	1.21	17.40
Freshwater fish	98.00	0.57	0.00	0.06	15.50	0.00	1.20	1.00	17.77
Marine fish	108.00	0.00	0.00	0.02	0.00	0.001	1.10	2.23	17.60
Shrimp/crab	93.50	1.15	0.00	0.05	0.00	0.00	3.45	2.96	16.95
Fresh milk	61.00	5.00	0.00	0.02	11.00	0.001	0.10	0.25	3.10
Powdered milk	504.00	39.00	0.00	0.00	0.00	0.00	0.00	0.00	24.00
Yoghurt	88.00	11.90	0.00	0.03	11.30	0.001	1.60	0.63	3.00
Egg	151.00	0.10	0.00	0.04	0.00	0.00	1.70	0.00	12.10
Soybean milk	30.00	1.20	0.00	0.02	5.00	0.001	0.40	0.28	3.00
Tofu	280.42	6.01	0.00	0.08	32.43	0.00	5.13	2.39	28.33
Green vegetables	18.33	2.50	0.00	0.03	103.90	25.00	2.10	0.50	1.93
Cabbage	9.00	2.60	0.00	0.02	5.30	11.00	0.20	0.23	1.10
Celery	11.00	3.10	1.00	0.01	13.60	2.00	0.20	0.14	0.40
Green beans	22.50	5.30	0.00	0.04	27.70	18.00	1.05	0.44	2.25
Carrots	25.00	8.10	0.00	0.00	4.80	9.00	0.30	0.22	1.00
Tomato	11.00	3.30	0.00	0.02	5.60	14.00	0.20	0.12	0.90
Eggplant	13.00	4.80	0.00	0.03	6.30	0.00	0.50	0.20	1.10
Potato	79.00	17.80	1.10	0.10	12.40	14.00	0.40	0.30	2.60
Mushroom	19.00	1.90	0.00	0.00	0.00	1.00	0.30	0.66	2.20
Pepper	24.50	11.45	11.80	0.09	20.55	72.50	0.45	0.27	2.45
Bamboo	22.00	3.00	0.00	0.07	0.00	8.50	0.50	0.31	2.15
Fungus	205.00	35.70	0.00	0.17	0.00	0.00	97.40	3.18	12.10
Seaweed	78.75	10.88	0.00	0.08	0.00	2.00	15.95	0.99	7.70
Garlic	126.00	26.50	0.00	0.04	0.00	7.00	1.20	0.88	4.50
Pickles	28.00	3.98	0.00	0.04	0.00	3.67	5.15	0.94	2.38
Nuts	457.83	19.08	9.17	0.25	152.80	10.00	4.84	3.93	20.88
Fruits	68.28	14.45	0.00	0.05	0.00	20.45	1.74	0.32	1.62

Table 3 Hierarchical Clustering Outcome Summary

Analysis of Variance for Variable bw Classified by Variable CLUSTER		
CLUSTER	**N**	**Mean**
3	2218	3267.80207
5	788	3276.25635
6	1553	3276.59433
1	2250	3307.38089
2	378	3265.78042
4	170	3239.70588

Table 4 K-means Clustering Outcome Summary

Cluster	Frequency	Mean	RMS Std Deviation	Distance Between Cluster Centroids
1	174	3246.26	0.6000	12.6513
2	432	3315.66	0.5177	4.6464
3	14	3307.14	4.9748	35.0998
4	3033	3263.17	0.2448	2.5642
5	11	3345.45	20.8868	71.4222
6	3712	3295.38	0.2011	2.5642

Table 5 P-values of Global Test Results in Subgroups

	Low BMI	Normal BMI	High BMI
Preconception	0.0439	0.3075	0.5186
1^{st} trimester	0.0033	0.2603	0.3883
2^{nd} trimester	0.0017	0.5784	0.3887
3^{rd} trimester	0.0024	0.2475	0.4624

Table 6 Base Model and Full Models with Principle Components for Each Time Period

	Base Model	Full Model (pre-preg)	Full Model (1st tri)	Full Model (2nd tri)	Full Model (3rd tri)
Intercept	-3889.70 (<0.0001)	-3786.22 (<0.0001)	-3767.00 (<0.0001)	-3761.38 (<0.0001)	-3778.00 (<0.0001)
Education level					
High School	10.99 (0.7416)	8.29 (0.8032)	7.11 (0.8310)	8.78 (0.7922)	6.46 (0.8464)
Community College	30.58 (0.3345)	23.78 (0.4525)	21.54 (0.4976)	22.80 (0.4732)	20.56 (0.5176)
College	30.68 (0.3433)	26.18 (0.4177)	25.35 (0.4338)	22.87 (0.4809)	22.30 (0.4920)
Graduate School	58.82 (0.2385)	54.97 (0.2694)	54.24 (0.2762)	56.15 (0.2601)	54.16 (0.2775)
Income (CNY/month)					
2,000-3,000	0.5836 (0.9827)	-0.53 (0.9844)	0.09 (0.9974)	-1.39 (0.9589)	-1.28 (0.9623)
3,000-4,000	47.85 (0.1132)	44.07 (0.1466)	47.36 (0.1180)	44.45 (0.1426)	47.14 (0.1199)
More than 4,000	38.51 (0.2062)	39.44 (0.1963)	44.45 (0.1450)	44.61 (0.1444)	46.88 (0.1252)
Smoke	-25.41 (0.3025)	-26.80 (0.2762)	-29.05 (0.2381)	-27.92 (0.2568)	-28.97 (0.2397)
Hypertension	-281.50 (<0.0001)	-282.95 (<0.0001)	-281.60 (<0.0001)	-295.17 (<0.0001)	-292.80 (<0.0001)
Nation-Han	15.71 (0.6677)	10.09 (0.7828)	13.15 (0.7190)	13.20 (0.7183)	17.23 (0.6377)
Diabetes	723.50 (0.0001)	738.22 (0.0001)	683.00 (0.0004)	679.68 (0.0004)	676.50 (0.0004)
Maternal age (years)	5.41 (0.0547)	5.73 (0.04166)	5.42 (0.0538)	5.51 (0.0498)	5.41 (0.0543)
Gestational age (weeks)	155.55 (<0.0001)	153.33 (<0.0001)	153.00 (<0.0001)	153.33 (<0.0001)	153.60 (<0.0001)
BMI	49.72 (<0.0001)	48.77 (0.0001)	48.68 (0.0001)	47.58 (<0.0001)	48.03 (0.0001)
Prin1		***-2.70 (0.0441)***	-1.31 (0.3304)	***-2.64 (0.0486)***	-2.24 (0.0959)
Prin2		-1.53 (0.6073)	1.69 (0.5533)	1.21 (0.6714)	0.00469 (0.9987)
Prin3		-0.65 (0.8610)	-0.07 (0.9849)	2.72 (0.4444)	2.72 (0.4487)
Prin4		5.81 (0.1682)	1.41 (0.7455)	7.47 (0.0931)	6.42 (0.1438)
Prin5		***14.44 (0.0080)***	10.73 (0.0533)	8.37 (0.1309)	9.86 (0.0827)
Prin6		11.21 (0.0928)	10.78 (0.1124)	2.56 (0.7004)	4.41 (0.5101)
Prin7		-10.67 (0.1467)	***-17.07 (0.0207)***	-13.44 (0.0655)	-11.11 (0.1220)
Prin8		4.22 (0.6204)	14.97 (0.0661)	-14.27 (0.0870)	12.11 (0.1384)
Prin9		***-20.53 (0.0224)***	***-19.73 (0.0308)***	10.23 (0.2726)	16.05 (0.0861)
Prin10		-2.39 (0.7990)	-0.23 (0.9803)	***-19.17 (0.0431)***	***-19.69 (0.0363)***
R-squared	0.3996	0.4094	0.4084	0.4086	0.4083

Table 7 Base Model and Full Models Based on Stepwise Selection for Each Time Period

	Base Model	Full Model (pre-preg)	Full Model (1st tri)	Full Model (2nd tri)	Full Model (3rd tri)
Intercept	-3889.70 (<0.0001)	-3246.75 (<0.0001)	-3320.24 (<0.0001)	-3236.94 (<0.0001)	-3270.22 (<0.0001)
Education level					
High School	10.99 (0.7416)	9.92 (0.7644)	4.32 (0.8963)	10.75 (0.7455)	9.82 (0.7671)
Community College	30.58 (0.3345)	27.99 (0.3739)	21.04 (0.5051)	26.82 (0.3944)	25.83 (0.4128)
College	30.68 (0.3433)	31.16 (0.3326)	26.40 (0.4122)	30.87 (0.3371)	29.16 (0.3652)
Graduate School	58.82 (0.2385)	58.18 (0.2410)	53.64 (0.2798)	53.57 (0.2802)	51.08 (0.3041)
Income (CNY/month)					
2,000-3,000	0.5836 (0.9827)	-2.77 (0.9176)	-5.75 (0.8304)	-6.52 (0.8078)	-6.47 (0.8097)
3,000-4,000	47.85 (0.1132)	44.53 (0.1383)	41.27 (0.1707)	39.80 (0.1862)	41.46 (0.1686)
More than 4,000	38.51 (0.2062)	38.30 (0.2057)	38.64 (0.2025)	36.40 (0.2292)	36.96 (0.2226)
Smoke	-25.41 (0.3025)	-23.97 (0.3274)	-31.36 (0.2013)	-26.15 (0.2854)	-27.13 (0.2683)
Hypertension	-281.50 (<0.0001)	-277.67 (<0.0001)	-285.70 (<0.0001)	-285.37 (<0.0001)	-284.74 (<0.0001)
Nation-Han	15.71 (0.6677)	12.39 (0.7334)	11.87 (0.7445)	12.51 (0.7310)	13.45 (0.7119)
Diabetes	723.50 (0.0001)	706.08 (0.0002)	730.12 (0.0001)	731.92 (0.0001)	735.91 (<0.0001)
Maternal age (years)	5.41 (0.0547)	5.58 (0.0459)	5.60 (0.0454)	5.78 (0.0387)	5.86 (0.0365)
Gestational age (weeks)	155.55 (<0.0001)	153.13 (<0.0001)	154.13 (<0.0001)	153.89 (<0.0001)	154.15 (<0.0001)
BMI	49.72 (<0.0001)	48.92 (<0.0001)	47.64 (<0.0001)	45.93 (0.0002)	46.84 (0.0001)
T0n6		***7.10 (0.0022)***			
T0n10		***-149.29 (<0.0001)***			
T1n16			***0.8327 (0.0015)***		
T1n14			***-0.5148 (<0.0001)***		
T2n6				***6.95 (0.0051)***	
T2n14				***-0.52 (<0.0001)***	
T3n6					***6.94 (0.0062)***
T3n14					***-0.48 (<0.0001)***
R-squared	0.3996	0.4085	0.4076	0.4087	0.4072

T0: pre-pregnancy; T1: 1st Trimester; T2: 2nd Trimester; T3: 3rd Trimester

n6: Insoluble dietary fiber, g; n10: Thiamin, mg; n14: Folate, ug; n16: Vitamin C, mg

Table 8 Base Model and Full Models Based on Lasso Method for Each Time Period

	Base Model	Full Model (pre-preg)	Full Model (1st tri)	Full Model (2nd tri)	Full Model (3rd tri)
Intercept	-3889.70 (<0.0001)	-3174.83 (<0.0001)	-3254.24 (<0.0001)	-3185.17 (0.0001)	-3220.10 (<0.0001)
Education level					
High School	10.99 (0.7416)	7.56 (0.8193)	7.47 (0.8222)	10.51 (0.7508)	8.89 (0.2984)
Community College	30.58 (0.3345)	25.34 (0.4206)	26.54 (0.4014)	26.86 (0.3935)	25.91 (0.4111)
College	30.68 (0.3433)	29.35 (0.3615)	28.94 (0.3700)	30.37 (0.3447)	28.95 (0.3684)
Graduate School	58.82 (0.2385)	52.92 (0.2860)	60.42 (0.2248)	54.18 (0.2744)	51.65 (0.2984)
Income (CNY/month)					
2,000-3,000	0.5836 (0.9827)	-5.92 (0.8249)	-5.64 (0.8343)	-5.59 (0.8348)	-6.18 (0.8180)
3,000-4,000	47.85 (0.1132)	38.42 (0.2020)	40.96 (0.1752)	41.16 (0.1714)	42.65 (0.1566)
More than 4,000	38.51 (0.2062)	36.06 (0.2332)	35.77 (0.2393)	37.34 (0.2172)	37.62 (0.2143)
Smoke	-25.41 (0.3025)	-26.04 (0.2872)	-25.81 (0.2937)	-24.33 (0.3203)	-25.09 (0.3063)
Hypertension	-281.50 (<0.0001)	-283.92 (<0.0001)	-282.67 (<0.0001)	-284.60 (<0.0001)	-284.60 (<0.0001)
Nation-Han	15.71 (0.6677)	9.15 (0.8013)	11.23 (0.7585)	13.32 (0.7141)	14.13 (0.6978)
Diabetes	723.50 (0.0001)	722.61 (0.0001)	725.10 (0.0001)	722.39 (0.0001)	727.43 (0.0001)
Maternal age (years)	5.41 (0.0547)	6.12 (0.0291)	5.61 (0.0459)	5.62 (0.0444)	5.71 (0.0416)
Gestational age (weeks)	155.55 (<0.0001)	152.94 (<0.0001)	154.77 (<0.0001)	153.14 (<0.0001)	153.45 (<0.0001)
BMI	49.72 (<0.0001)	46.85 (0.0001)	47.44 (0.0001)	46.45 (0.0001)	47.13 (0.0001)
T0n6		***8.62 (0.0004)***			
T0n10		-78.60 (0.1160)			
T0n14		-0.25 (0.0876)			
T0n37		-0.03 (0.3187)			
T1n10			-41.85 (0.2468)		
T1n14			-0.23 (0.0716)		
T2n6				***7.73 (0.0021)***	
T2n10				-58.08 (0.0813)	
T2n14				***-0.42 (0.0010)***	
T3n6					***7.64 (0.0030)***
T3n10					-52.63 (0.1055)
T3n14					***-0.3776 (0.0036)***
R-square	0.3996	0.4052	0.4039	0.4100	0.4083

T0: pre-pregnancy; T1: 1st Trimester; T2: 2nd Trimester; T3: 3rd Trimester
n6: Insoluble dietary fiber, g; n10: Thiamin, mg; n14: Folate, ug; n37: Sulfur-containing ami-no acid(SAA), mg.

Figures

Figure 1 Geographic Information of Lanzhou, China

Figure 2 Histograms of Birth Weight and Continuous Confounder Variables

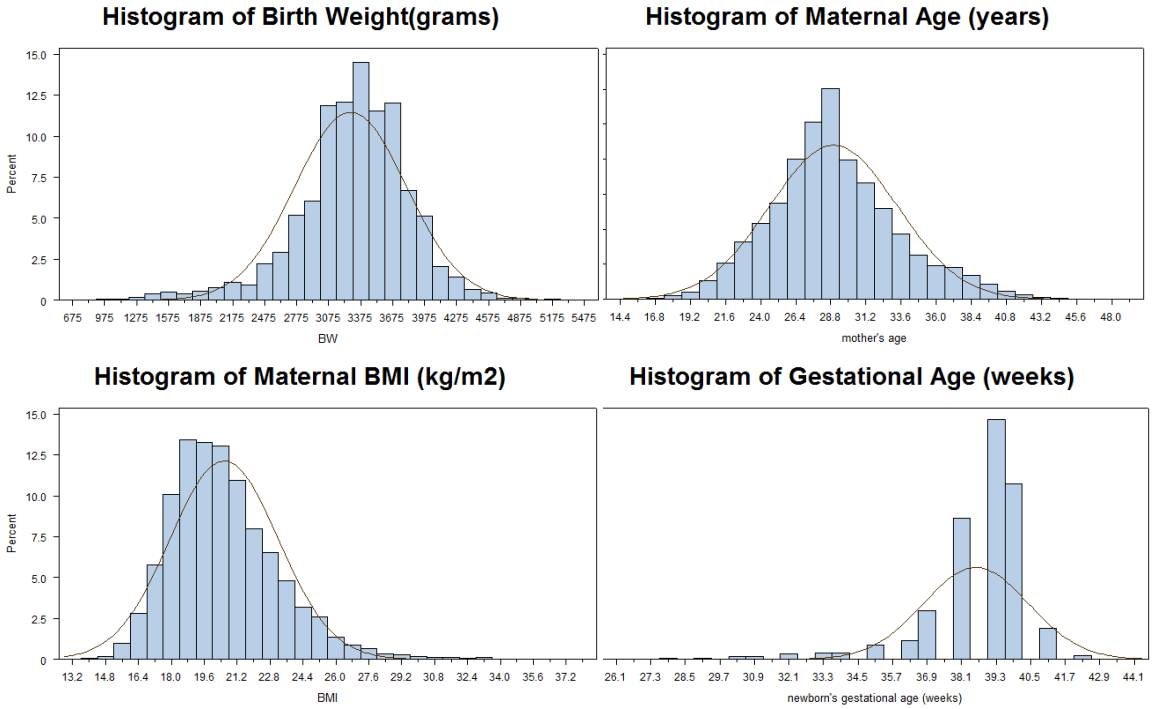


Figure 3 Food Intake against Birth Weight

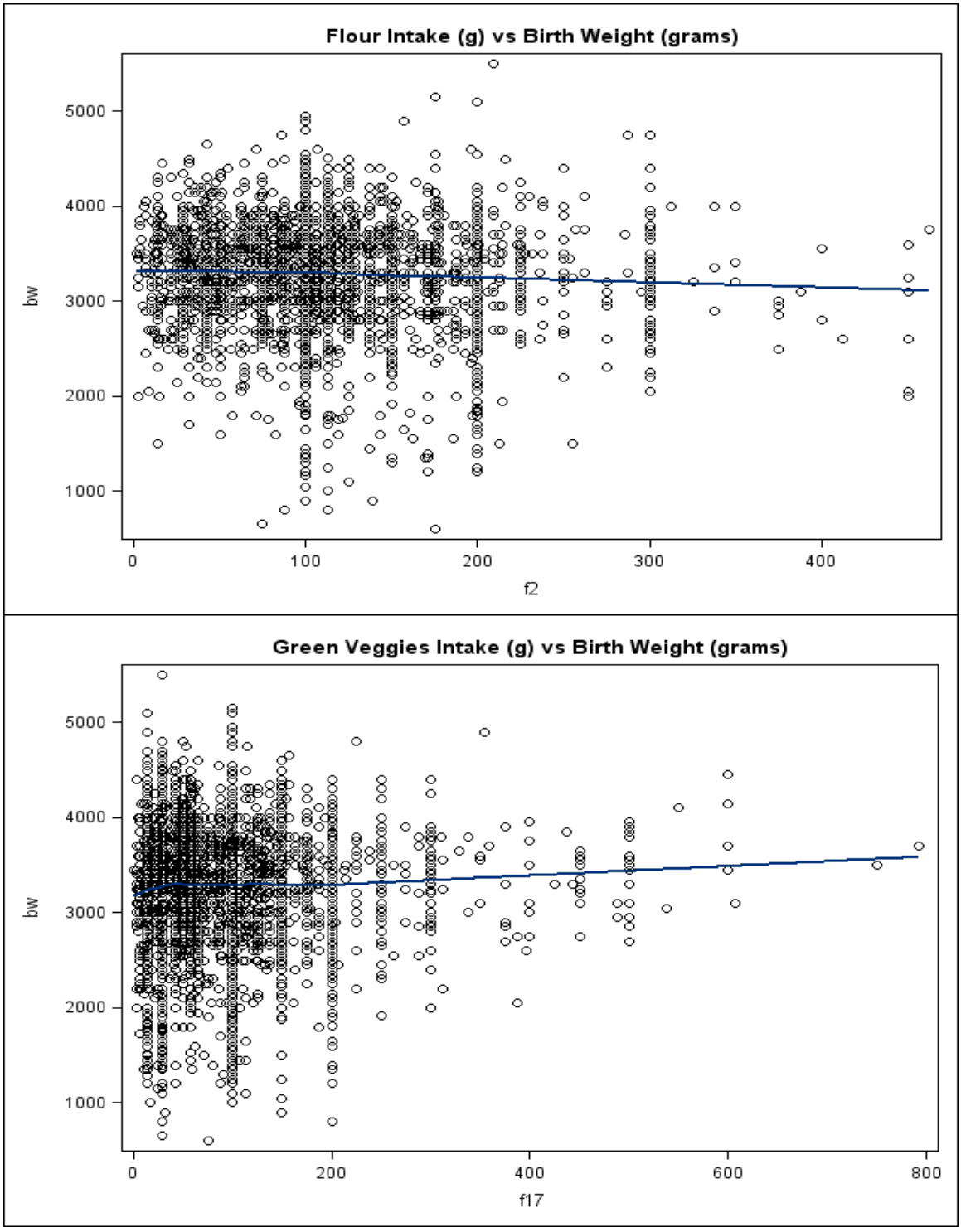
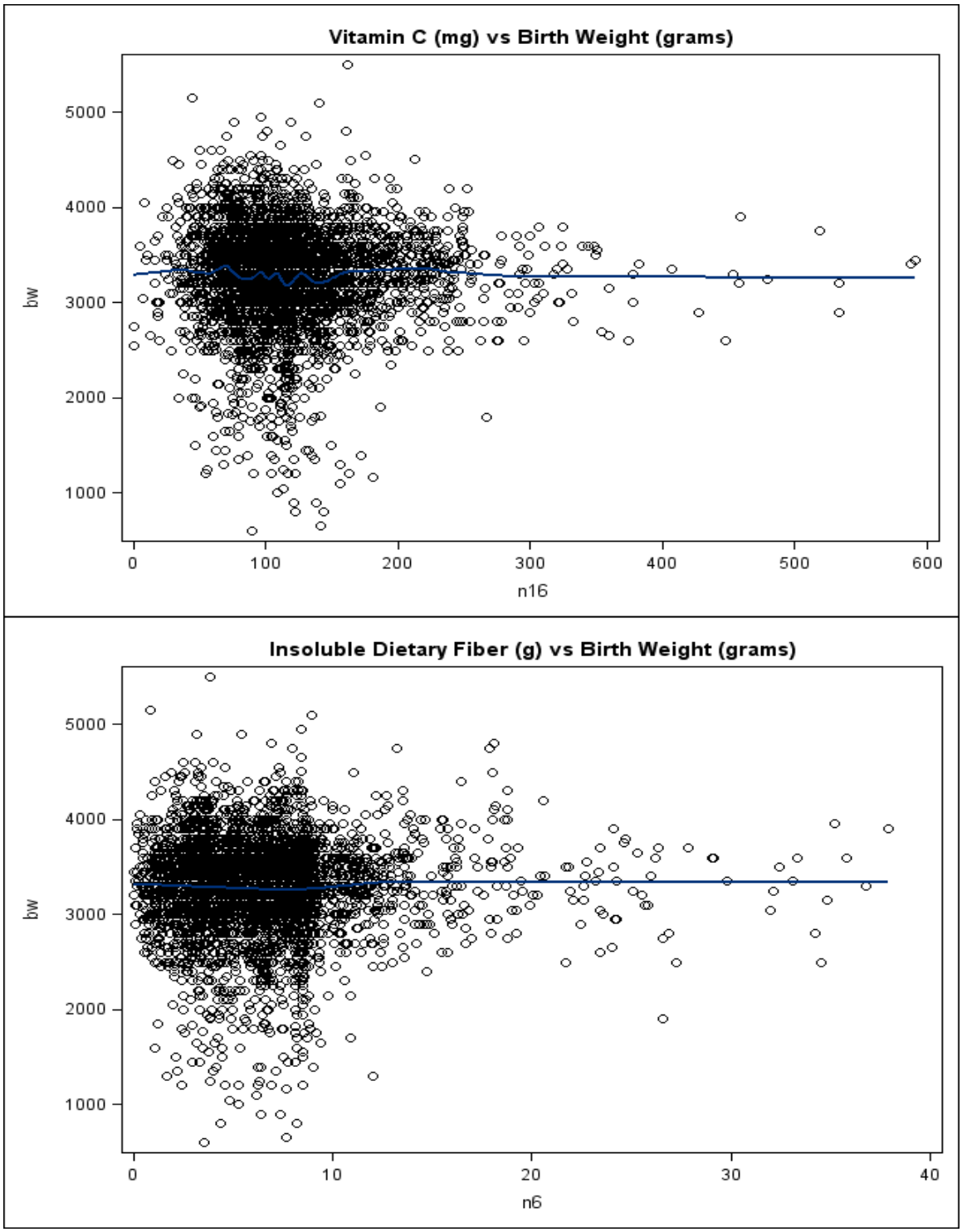


Figure 4 Nutrient Intake against Birth Weight



ACKNOWLEDGEMENTS

There are many people that I want to express my gratitude to for helping me finish this thesis and accomplish my two-year study at Yale. First of all, I really appreciate all the guidance and encouragement from my academic and thesis advisor, Prof. Shuangge Ma. From the very beginning of my study at Biostatistics to the finishing of this thesis, he has taught me a lot during classes and invaluable discussion meetings with his precious time. I would also like to thank my second reader, Prof. Yawei Zhang, who offered me this great opportunity of getting involved in this study and helped me with the understanding of the nutrition background of my thesis. I am thankful to all the professors at Department of Biostatistics and Yale School of Public Health. Last, I would like to thank my family and friends for their support and help during these years.

REFERENCES

Akcakus, M.*, et al.* (2006) The relationship between birthweight, 25-hydroxyvitamin D concentrations and bone mineral status in neonates, *Annals of tropical paediatrics*, 26, 267-275.

Almeida, J.A.S.*, et al.* (2007) Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering, *Chemometrics and Intelligent Laboratory Systems*, 87, 208-217.

Atallah, A.N., Hofmeyr, G.J. and Duley, L. (2002) Calcium supplementation during pregnancy for preventing hypertensive disorders and related problems, *Cochrane Database Syst Rev*, CD001059.

Barker, M.*, et al.* (1997) Birth weight and body fat distribution in adolescent girls, *Archives of disease in childhood*, 77, 381-383.

Bhutta, Z.A.*, et al.* (2009) A comparative evaluation of multiple micronutrient and iron-folic acid supplementation during pregnancy in Pakistan: impact on pregnancy outcomes, *Food and nutrition bulletin*, 30, S496-505.

Bock, H.H. (1985) On Some Significance Tests in Cluster-Analysis, *J Classif*, 2, 77-108.

Bodnar, L.M.*, et al.* (2010) Maternal serum 25-hydroxyvitamin D concentrations are associated with small-for-gestational age births in white women, *The Journal of nutrition*, 140, 999-1006.

Bonellie, S.R. (2012) Use of multiple linear regression and logistic regression models to investigate changes in birthweight for term singleton infants in Scotland, *Journal of clinical nursing*, 21, 2780-2788.

Bouwland-Both, M.*, et al.* (2013) A periconceptional energy-rich dietary pattern is associated with early fetal growth: the Generation R study, *BJOG : an international journal of obstetrics and gynaecology*, 120, 435-445.

Campbell, M.K.*, et al.* (2012) Determinants of small for gestational age birth at term, *Paediatric and perinatal epidemiology*, 26, 525-533.

Catov, J.M.*, et al.* (2011) Periconceptional multivitamin use and risk of preterm or small-for-gestational-age births in the Danish National Birth Cohort, *The American journal of clinical nutrition*, 94, 906-912.

Choo, V. (2002) WHO reassesses appropriate body-mass index for Asian populations, *The Lancet*, 360, 235.

Curhan, G.C.*, et al.* (1996) Birth weight and adult hypertension, diabetes mellitus, and obesity in US men, *Circulation*, 94, 3246-3250.

Donahue, S.M.*, et al.* (2010) Trends in birth weight and gestational length among singleton term births in the United States: 1990-2005, *Obstetrics and gynecology*, 115, 357-364.

Dror, D.K. and Allen, L.H. (2012) Interventions with vitamins B6, B12 and C in pregnancy, *Paediatric and perinatal epidemiology*, 26 Suppl 1, 55-74.

Efron, B.*, et al.* (2004) Least angle regression, *Ann Stat*, 32, 407-451.

Fall, C.H.*, et al.* (2009) Multiple micronutrient supplementation during pregnancy in low-income countries: a meta-analysis of effects on birth size and length of gestation, *Food and nutrition bulletin*, 30, S533-546.

Fuster, V.*, et al.* (2013) Factors determining the variation in birth weight in Spain (1980-2010), *Annals of human biology*.

Gale, C.R.*, et al.* (2008) Maternal vitamin D status during pregnancy and child outcomes, *European journal of clinical nutrition*, 62, 68-77.

Gebremedhin, S., Enquselassie, F. and Umeta, M. (2012) Independent and joint effects of prenatal Zinc and Vitamin A Deficiencies on birthweight in rural Sidama, Southern Ethiopia: prospective cohort study, *PloS one*, 7, e50213.

Goeman, J.J., et al. (2003) A global test for groups of genes: testing association with a clinical outcome, Bioinformatics, 20, 93-99.

Gorodischer, R.*, et al.* (1995) Differences in cord serum retinol concentrations by ethnic origin in the Negev (southern Israel), *Early human development*, 42, 123-130.

Guttman, L. (1954) Some Necessary Conditions for Common-Factor Analysis, *Psychometrika*, 19, 149-161.

Han, Z.*, et al.* (2011) Maternal underweight and the risk of preterm birth and low birth weight: a systematic review and meta-analyses, *International journal of epidemiology*, 40, 65-101.

Hartigan, J.A. (1985) Statistical-Theory in Clustering, *J Classif*, 2, 63-76.

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components, *J Educ Psychol*, 24, 498-520.

Hunt, K.J.*, et al.* (2013) Maternal pre-pregnancy weight and gestational weight gain and their association with birthweight with a focus on racial differences, *Maternal and child health journal*, 17, 85-94.

Johnson, S.C. (1967) Hierarchical Clustering Schemes, *Psychometrika*, 32, 241-254.

Kaiser, H.F. and Caffrey, J. (1965) Alpha Factor Analysis, *Psychometrika*, 30, 1-14.

Kelly, A.*, et al.* (1996) A WHO Collaborative Study of Maternal Anthropometry and Pregnancy Outcomes, *International journal of gynaecology and obstetrics: the official organ of the International Federation of Gynaecology and Obstetrics*, 53, 219-233.

Kramer, M.S. (1987) Determinants of low birth weight: methodological assessment and meta-analysis, *Bulletin of the World Health Organization*, 65, 663-737.

Kramer, M.S. and Kakuma, R. (2003) Energy and protein intake in pregnancy, *Cochrane Database Syst Rev*, CD000032.

Kramer, M.S.*, et al.* (2000) Socio-economic disparities in pregnancy outcome: why do the poor fare so poorly?, *Paediatric and perinatal epidemiology*, 14, 194-210.

Lechtig, A.*, et al.* (1975) Influence of maternal nutrition on birth weight, *The American journal of clinical nutrition*, 28, 1223-1233.

Mahomed, K. (2000) Iron supplementation in pregnancy, *Cochrane Database Syst Rev*, CD000117.

Mahomed, K., Bhutta, Z. and Middleton, P. (2007) Zinc supplementation for improving pregnancy and infant outcome, *Cochrane Database Syst Rev*, CD000230.

Malina, R.M., Katzmarzyk, P.T. and Beunen, G. (1996) Birth weight and its relationship to size attained and relative fat distribution at 7 to 12 years of age, *Obesity research*, 4, 385-390.

Matte, T.D.*, et al.* (2001) Influence of variation in birth weight within normal range and within sibships on IQ at age 7 years: cohort study, *Brit Med J*, 323, 310-314.

McCance, D.R.*, et al.* (1994) Birth weight and non-insulin dependent diabetes: thrifty genotype, thrifty phenotype, or surviving small baby genotype?, *BMJ*, 308, 942-945.

Merialdi, M.*, et al.* (2003) Nutritional interventions during pregnancy for the prevention or treatment of impaired fetal growth: an overview of randomized controlled trials, *The Journal of nutrition*, 133, 1626S-1631S.

Michels, K.B.*, et al.* (1996) Birthweight as a risk factor for breast cancer, *Lancet*, 348, 1542-1546.

Milligan, G.W. and Cooper, M.C. (1985) An Examination of Procedures for Determining the Number of Clusters in a Data Set, *Psychometrika*, 50, 159-179.

Milligan, G.W. and Cooper, M.C. (1987) Methodology Review - Clustering Methods, *Appl Psych Meas*, 11, 329-354.

Mohsin, M.*, et al.* (2003) Maternal and neonatal factors influencing premature birth and low birth weight in Australia, *Journal of biosocial science*, 35, 161-174.

Mori, R.*, et al.* (2012) Zinc supplementation for improving pregnancy and infant outcome, *Cochrane Database Syst Rev*, 7, CD000230.

Morley, R.*, et al.* (2009) Maternal 25-hydroxyvitamin D concentration and offspring birth size: effect modification by infant VDR genotype, *European journal of clinical nutrition*, 63, 802-804.

Osborne, M.R., Presnell, B. and Turlach, B.A. (2000) On the LASSO and its dual, *J Comput Graph Stat*, 9, 319-337.

Osrin, D.*, et al.* (2005) Effects of antenatal multiple micronutrient supplementation on birthweight and gestational duration in Nepal: double-blind, randomised controlled trial, *Lancet*, 365, 955-962.

Ounsted, M.K., Moar, V.A. and Scott, A. (1984) Children of deviant birthweight at the age of seven years: health, handicap, size and developmental status, *Early human development*, 9, 323-340.

Parker, J.D., Schoendorf, K.C. and Kiely, J.L. (1994) Associations between measures of socioeconomic status and low birth weight, small for gestational age, and premature delivery in the United States, *Annals of epidemiology*, 4, 271-278.

Pollard, D. (1981) Strong Consistency of K-Means Clustering, *Ann Stat*, 9, 135-140.

Prentice, A.*, et al.* (2009) Maternal plasma 25-hydroxyvitamin D concentration and birthweight, growth and bone mineral accretion of Gambian infants, *Acta Paediatr*, 98, 1360-1362.

Record, R.G., McKeown, T. and Edwards, J.H. (1969) The relation of measured intelligence to birth weight and duration of gestation, *Annals of human genetics*, 33, 71-79.

Rich-Edwards, J.W.*, et al.* (1999) Birthweight and the risk for type 2 diabetes mellitus in adult women, *Annals of internal medicine*, 130, 278-284.

Rich-Edwards, J.W.*, et al.* (1997) Birth weight and risk of cardiovascular disease in a cohort of women followed up since 1976, *BMJ*, 315, 396-400.

Richards, M.*, et al.* (2002) Birthweight, postnatal growth and cognitive function in a national UK birth cohort, *International journal of epidemiology*, 31, 342-348.

Rumbold, A. and Crowther, C.A. (2005) Vitamin C supplementation in pregnancy, *Cochrane Database Syst Rev*, CD004072.

Sarle, W.S. (1983) Cubic Clustering Criterion. *SAS Technical Report A-108, Cary, NC: SAS Institute Inc.*

Singhal, A.*, et al.* (2003) Programming of lean body mass: a link between birth weight, obesity, and cardiovascular disease?, *The American journal of clinical nutrition*, 77, 726-730.

Susser, M. and Stein, Z. (1982) Third variable analysis: application to causal sequences among nutrient intake, maternal weight, birthweight, placental weight, and gestation, *Statistics in medicine*, 1, 105-120.

Symons, M.J. (1981) Clustering Criteria and Multivariate Normal Mixtures, *Biometrics*, 37, 35-43.

Szekely, G.J. and Rizzo, M.L. (2005) Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method, *J Classif*, 22, 151-183.

Tamura, T.*, et al.* (1997) Serum concentrations of zinc, folate, vitamins A and E, and proteins, and their relationships to pregnancy outcome, *Acta obstetricia et gynecologica Scandinavica. Supplement*, 165, 63-70.

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso, *J Roy Stat Soc B Met*, 58, 267-288.

Timmermans, S.*, et al.* (2012) The Mediterranean diet and fetal size parameters: the Generation R Study, *The British journal of nutrition*, 108, 1399-1409.

Tsimbos, C. and Verropoulou, G. (2011) Demographic and socioeconomic determinants of low birth weight and preterm births among natives and immigrants in Greece: an analysis using nationwide vital registration micro-data, *Journal of biosocial science*, 43, 271-283.

van den Broek, N.*, et al.* (2010) Vitamin A supplementation during pregnancy for maternal and newborn outcomes, *Cochrane Database Syst Rev*, CD008666.

Ward, J.H. (1963) Hierarchical Grouping to Optimize an Objective Function, *J Am Stat Assoc*, 58, 236-&.

Wells, J.C. (2002) Thermal environment and human birth weight, *Journal of theoretical biology*, 214, 413-425.

Zagre, N.M.*, et al.* (2007) Prenatal multiple micronutrient supplementation has greater impact on birthweight than supplementation with iron and folic acid: a cluster-randomized, double-blind, controlled programmatic study in rural Niger, *Food and nutrition bulletin*, 28, 317-327.

Zeng, L.*, et al.* (2008) Impact of micronutrient supplementation during pregnancy on birth weight, duration of gestation, and perinatal mortality in rural western China: double blind cluster randomised controlled trial, *BMJ*, 337, a2001.

Zhang, J. and Bowes, W.A., Jr. (1995) Birth-weight-for-gestational-age patterns by race, sex, and parity in the United States population, *Obstetrics and gynecology*, 86, 200-208.