Appendix U Statistical methods for the descriptive statistics, comparison of dietary intake and trends over time in Years 1 to 11 (2008/09 – 2018/19) of the NDNS Rolling Programme (RP)

U.1 Introduction

This appendix provides an outline description of the statistical methods used for the following:

- descriptive statistics used in this report
- comparisons of dietary intake for non-overlapping subpopulations, defined by grouped fieldwork years, from Years 1 to 11 of the NDNS RP
- assessment of trends over time from Years 1 to 11 (2008/09 – 2018/19)

The NDNS RP sample requires weights to adjust for differences in sample selection and response relative to the UK population distribution. The statistical analysis of data generated from this complex survey design requires taking the sample design (i.e. sample stratification, clustering and weighting) into account to yield valid estimates of the population parameters. Details of the weighting and sampling procedures are provided in appendix B.

U.2 Descriptive statistics used in this report

The choice of descriptive statistic is mainly driven by the statistical distribution of the data for each variable:

- A numerical variable which follows a symmetric and 'bell-shaped' distribution is best described using an arithmetic mean (to represent the typical value) and standard deviation (to represent the spread). This report also provides the median and 2.5th and 97.5th percentiles which provide robust (not outlier influenced) estimates of the typical value and spread of the distribution for the

case when the numerical variable deviates from a symmetric and 'bell-shaped' distribution due to extreme outliers or a high proportion of zeros

- A numerical variable which is positively skewed (bunched for low values and widely spread for high values) is best described using a geometric mean (to represent the typical value) and 2.5th and 97.5th percentiles (to represent the spread) as the arithmetic mean will be strongly influenced by the relatively few high outlier values
- A numerical variable which has a high proportion of values below the limit of quantitation[i] is best described using a median (to represent the typical value) and 2.5th and 97.5th percentiles (to represent the spread)

Evidence from literature is used to confirm the choice of descriptive statistic for each variable before it is used in this report.

U.2.1 Descriptive statistics used for food, nutrient intake and blood and urine analyte variables

The majority of food, nutrient intake and blood and urine analyte variables reported follow a symmetric and 'bell-shaped' distribution and so the descriptive statistics used are arithmetic mean, median, standard deviation, 2.5th and 97.5th percentiles. Exceptions to this are outlined in the table below along with the reported descriptive statistics:

	Distribution	**Descriptive statistics**	
Blood analyte: Red cell blood folate	Positively skewed	Geometric mean	2.5th and 97.5th percentiles
Blood analyte: Serum folate	Positively skewed	Geometric mean	2.5th and 97.5th percentiles
Blood analyte: Unmetabolised (free) folic acid	High proportion of values below the limit of quantitation	Median	2.5th and 97.5th percentiles
Urine analyte: Iodine concentration	Positively skewed (but following WHO guidance) [ii]	Median	20th and 80th percentiles

[i] The limit of quantitation is the lowest amount that can reliably and consistently be detected and measured.

[ii] http://whqlibdoc.who.int/publications/2007/9789241595827_eng.pdf

The variables listed below show some evidence of deviating from a symmetric and 'bell-shaped' distribution and so the more robust (not outlier influenced) median and 2.5th and 97.5th percentiles should be used for interpretation rather than the arithmetic mean and standard deviation.

Food intake: Sugar-sweetened soft drinks, Sugar confectionery, Chocolate confectionery.
Nutrient intake: Vitamin A, Vitamin D.
Blood analytes: Ferritin, Vitamin B6 (PLP) and Vitamin B12.

U.3 Comparison of dietary intake between non-overlapping subpopulations

This section outlines the statistical methods used to estimate the differences between mean intakes of key foods and nutrients from non-overlapping subpopulations for grouped fieldwork years. NDNS RP data for Years 1 to 11 were split to form 5 groups (survey period 1: Years 1 and 2, survey period 2: Years 3 and 4, survey period 3: Years 5 and 6, survey period 4: Years 7 and 8 and survey period 5: Years 9, 10 and 11).

The same weights and design variables created for the Years 1 to 11 dataset were applied to the appropriate subsets of the data.[iii] Analysis of mean daily intake of key nutrients and foods compared Years 7 and 8 (combined) with Years 9, 10 and 11 (combined) across 7 age groups, overall and by sex (for all but the youngest age group). The age groups were 1.5 to 3 years (sex-combined only), 4 to 10 years, 11 to 18 years, 19 to 64 years, 65 years and over and then additionally 65 to 74 years and 75 years and over. Previous NDNS reports have compared Years 7 and 8

[iii] Although the weights were not specifically designed for this type of sub-group analysis, it was possible to use the Years 1 to 11 weights and design variables for just 2 or 3 years' data (Years 1 and 2, Years 3 and 4, Year 5 and 6, Years 7 and 8 or Years 9, 10 and 11), as:

- the selection weights correct for any differences in sampling strategy across survey years
- there was no evidence that response behaviour had changed significantly between the 5 survey periods

However, to use subsets of any other combination of years of the dataset, the weights and design variables would have to be reviewed to ensure that the subset of data is still representative of the UK population when the Years 1 to 11 weights and design variables have been applied.

(combined) with Years 1 and 2 (combined), Years 5 and 6 (combined) with Years 1 and 2 (combined) and Years 3 and 4 (combined) with Years 1 and 2 (combined).[1]

The comparisons described above involve comparing either means of continuous variables (mean differences in energy and nutrient intakes) or differences of proportions (such as the percentage of the sample meeting the 5 A Day guideline for fruit and vegetable intake[iv]) among groups, defined by survey periods (Years 7 and 8 (combined) compared with Years 9, 10 and 11 (combined)) overall and by sex. The mean differences for the continuous variables were estimated through linear regression models and differences of proportions through logistic regression models. Each regression model included all 5 survey periods and comparisons of interest were selected from this model. The statistical analyses were undertaken following 3 stages: exploratory analyses, estimation of mean differences and diagnostic procedures (i.e. assessment of model assumptions and goodness of fit). All the analyses, including the graphical tools and diagnostic procedures, took into account the complex survey design.

U.3.1 Exploratory analyses

The observed distribution of the continuous variables was screened through histograms, Q-Q plots and boxplots. These graphical tools showed the shape of the distribution and highlighted the presence of outliers. These were investigated as well as their impact on the regression analyses. In cases where the variable had small variability and hence took a reduced range of values (e.g. red and processed meat), the variable was dichotomised using the population median as the cut-off value and analysed through logistic regression.

U.3.2 Estimation of differences of means

Linear regression models were used for continuous measurements of nutrient or food intake. The purpose of the analyses was to perform simple study-domain

[iv] Appendix A provides further details regarding the 5 A Day guidelines for those aged 11 years and over. 5 A Day portions of fruit and vegetables were not calculated for children aged 10 years and younger.

comparisons rather than investigating the relationship between nutrient or food intake and age or sex. Therefore, only categorical variables needed to be defined to represent the comparison groups (Years 7 and 8 (combined) compared with Years 9, 10 and 11 (combined)), the study domains (age and sex) and their interactions. The regression coefficients estimate the subgroup differences that exist in the population. This approach is equivalent to estimating each difference of means by study domain, provided that the full sample is used for the estimation of standard errors. The use of regression models allows the analyst to estimate the mean differences simultaneously. For illustration, consider the comparison of mean intakes of red and processed meat in grams between survey period 3 (Years 5 and 6) and survey period 1 (Years 1 and 2) across age groups. The response variable is red and processed meat intake and the independent variables are: age (categorical variable for 1.5 to 3 years, 4 to 10 years, 11 to 18 years, 19 to 64 years and 65 years and over), survey period (categorical variable for survey periods 3 and 1) and the interaction between age and survey period. The variable “age” has 4 associated regression coefficients (B11, B12, B13 and B14), the indicator variable “survey period” has 1 regression coefficient (B2), the interaction term generates 4 regression coefficients (B31, B32, B33 and B34), and the intercept is denoted by B0. The target differences of means are functions of these parameters as described in table U.1. Tests of hypothesis for these differences can be undertaken by use of the estimated regression parameters and their covariance matrix.

Table U.1 Comparison of mean intakes of red and processed meat in grams between survey periods 3 and 1 across age groups in terms of linear regression parameters

Age group (years)	Mean intake (survey period 1)	Mean intake (survey period 3)	Difference of means (survey period 3 minus period 1)
1.5-3	B0	B0+B2	B2
4-10	B0+B11	B0+B11+B2+B31	B2+B31
11-18	B0+B12	B0+B12+B2+B32	B2+B32
19-64	B0+B13	B0+B13+B2+B33	B2+B33
65 years and over	B0+B14	B0+B14+B2+B34	B2+B34

In this example the linear regression model can be expressed as:

$$y_{hij} = B0 + \sum_{r=1}^{4} B1rx1r_{hij} + B2x3_{hij} + \sum_{r=1}^{4} B3rx1r_{hij} \cdot x3_{hij} + \varepsilon_{hij}$$

where y_{hij} represents the observed red and processed meat intake for the *j*-th individual in the *i*-th primary sampling unit of the *h*-th stratum; x1r (r=1,2,3,4) are indicators for age groups, with the first group used as reference category; x3 is an indicator for survey period 3 and ε_{hij} is the error term.

The regression coefficients in this model were estimated using probability weighted least squares[2] and their covariance matrix was estimated using a Taylor linearization method. [3]

U.3.3 Estimation of differences of proportions

Logistic regression models (with an identity link function) were used for binary variables. The regression coefficients (which estimate the proportion parameters for each age/sex group) use a pseudo-likelihood approach**Error! Bookmark not defined.** and their covariance matrix was estimated using a Taylor linearization method.[3] The proportion parameter (along with the associated 95% confidence interval) estimates the average change in the proportion of people above the threshold for each variable.

U.3.4 Diagnostic procedures

The linearity assumption between the dependent variable and the explanatory variables is crucial in regression analyses; however, the use of categorical variables as independent explanatory variables does not require the assumption of a linear relationship with the dependent variable. Similarly, the logistic regressions specified above do not require a linear relationship between the proportion and the explanatory variables. Therefore, checks for departures from linearity were not undertaken. The goodness of fit of the linear models was examined using the concept of explained variation (R-squared).

U.4 Trends over time

This section outlines the statistical methods used to estimate the average change per year in each outcome for key foods, nutrients and blood analytes from Years 1 to 11 of the NDNS RP. The same weights and design variables as those used in the Years 1 to 4 (combined), Years 5 and 6 (combined) and Years 7 and 8 (combined) reports (with additional weights and design variables for Years 9, 10 and 11 (combined)) were applied in these analyses. The weights for each data set were re-scaled based on sample size, such that each set of data is in the correct proportion (4:2:2:3) to give a standardised sample size per survey year.[v]

The average change per year for the continuous variables were estimated through linear regression models and for proportions (such as the percentage of the sample meeting the 5 A Day guideline for fruit and vegetable intake) through logistic regression models across 7 age groups, overall and by sex (for all but the youngest age group). The age groups were 1.5 to 3 years (sex-combined only), 4 to 10 years, 11 to 18 years, 19 to 64 years, 65 years and over and then additionally 65 to 74 years and 75 years and over. Participants were grouped into quarters of a calendar year according to when they completed their diary or when their blood sample was collected, and this was used as the explanatory variable in the regression models.

The statistical analyses were undertaken using the following 3 stages: exploratory analyses, estimation of changes per year and diagnostic procedures (i.e. assessment of model assumptions and goodness of fit). All the analyses, including the graphical tools and diagnostic procedures, took into account the complex survey design.

---

[v] Although the weights were not specifically designed for this type of sub-group analysis, it was possible to use the Years 1 to 11 weights and design variables for just 2 or 3 years' data (Years 1 and 2, Years 3 and 4, Year 5 and 6, Years 7 and 8 or Years 9, 10 and 11), as:

- the selection weights correct for any differences in sampling strategy across survey years
- there was no evidence that response behaviour had changed significantly between the 5 survey periods

However, to use subsets of any other combination of years of the dataset, the weights and design variables would have to be reviewed to ensure that the subset of data is still representative of the UK population when the Years 1 to 11 weights and design variables have been applied.

U.4.1 Exploratory analyses

The observed distributions of the continuous variables were screened through histograms, Q-Q plots and boxplots. These graphical tools showed the shape of the distribution and highlighted the presence of outliers. These were investigated as well as their impact on the regression analyses.

U.4.2 Estimation of changes per year for continuous variables

Linear regression models were used for continuous measurements of foods, nutrients and blood analytes. The regression coefficients (which estimate the intercept and slope parameters for each age/sex group) use probability weighted least squares[2] and their covariance matrix was estimated using a Taylor linearization method.[3] The slope parameter (along with the associated 95% confidence interval) estimates the average change per year for each variable.

U.4.3 Estimation of changes per year for proportions

Logistic regression models (with an identity link function) were used for binary variables. The regression coefficients (which estimate the intercept and slope parameters for each age/sex group) use a pseudo-likelihood approach**Error! Bookmark not defined.** and their covariance matrix was estimated using a Taylor linearization method.[3] The slope parameter (along with the associated 95% confidence interval) estimates the average change per year for each variable.

U.4.4 Diagnostic procedures

The goodness of fit of the linear models was examined using the concept of explained variation (R-squared).

U.4.5 Calculation of 11-year average change

Calculation of the 11-year average change is slightly different according to whether variables are analysed on the linear or the log scale (dependent on whether the data is highly skewed).

For variables analysed on the linear scale,

- multiply the average change per year by 11 to get the average change over 11 years e.g.:
    - Average change per year = -0.2mg/day (a reduction of 0.2mg/day per year).

      Average change over 11 years is 11*-0.2 = -2.2mg/day

    - Average change per year = +0.2mg/day (an increase of 0.2mg/day per year).

      Average change over 11 years is 11*0.2 = +2.2mg/day

For variables analysed on the log scale,

- convert the average percent change per year into a ratio of geometric means (divide by 100 and add 1), multiply this by itself 11 times (i.e. calculate it to the power of 11) and then convert back to a percent change over the 11 years (subtract 1 and multiply by 100). This will give a different percent change depending on whether it is a reduction or increase per year e.g.:
    - Average percent change per year = -3% (a reduction of 3% per year). This is equivalent to a ratio of 0.97 between yearly geometric means

      Average %change over 11 years is ((0.97)^11-1)*100= -28%

    - Average percent change per year = +3% (an increase of 3% per year). This is equivalent to a ratio of 1.03 between yearly geometric means

      Average % change over 11 years is ((1.03)^11-1)*100 = +38%

Average 11-year changes for each outcome of key foods, nutrients and blood analytes are provided in Excel tables U.1-U.4.

U.5 General

The statistical analyses described above were performed using the survey package in the statistical program R.[4,5]

The statistical analyses described in this appendix are for descriptive purposes rather than analytical, i.e. they are not intended to estimate the associations among many variables. Therefore, corrections for multiple comparisons were not necessary. Bonferroni procedures may be applicable in other situations involving simultaneous testing of regression coefficients when the number of independent variables in the regression analysis is large compared to the number of sampled Primary Sampling Units (PSUs).[6]

References

[1] National Diet and Nutrition Survey. [Internet]. Available from: www.gov.uk/government/collections/national-diet-and-nutrition-survey#current-ndns-results
[2] Holt, D., Smith, T.M.F. and Winter, P.D. (1980) Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society A,* **143,** 474 –487.
[3] Binder, D. A. (1983) On the Variances of Asymptotically Normal Estimators from Complex Surveys. International Statistical Review 51: 279–292
[4] Lumley, T. (2012) "survey: analysis of complex survey samples". R package version 3.28-2.
Lumley, T. (2004) Analysis of complex survey samples. *Journal of Statistical Software,* **9**(1): 1-19
[5] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
[6] Korn, E.L., Graubard, B.I.(1990) Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni *t* statistics. *The American Statistician,* **44,** 270 –276.