Appendix U Statistical methods for National Diet and Nutrition Survey 2019 to 2023

U.1 Introduction

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

- usual intake estimation of nutrient and food intakes
- descriptive statistics used in this report
- assessment of trends over time from Years 1 to 15 (2008 to 2023)
- trends in relation to socio-economic status of the participant's household (using equivalised income and index of multiple deprivation (IMD)).

The National Diet and Nutrition Survey (NDNS) 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 (that is the 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 BB.

U.2 Definitions of statistical terms used in this report

97.5th percentile
97.5% of the data are below this value and 2.5% are above it.

2.5th percentile
2.5% of the data are below this value and 97.5% are above it.

Confidence intervals
The range of plausible values for the true population estimate. Values outside this interval are statistically significantly different from the sample estimate.

Positive and negative skew
A skewed distribution does not have a symmetric shape. Positive skew occurs when there is a long tail on the right (a wide spread of extreme high values compared to the majority of the data) and negative skew occurs when there is a long tail on the left (a wide spread of extreme low values compared to the majority of the data).

Transformation
The application of a function to each data point such that the transformed data more closely follow a desired distribution (i.e. log transformation for positively skew data).

Arithmetic mean (mean)
The average of a set of numerical values, as calculated by adding them together and dividing by the number of values.

Geometric mean
An alternative to the arithmetic mean to estimate the average value for data that are positively skewed (one alternative to not having a symmetric shape).

Median
The middle value in a dataset when the data is ranked by size

U.3 Usual intake estimation of nutrient and food intakes

This report is the first NDNS report to present dietary data collected using Intake24, an online 24-hour dietary recall tool. The move to non-consecutive day dietary recalls for years 12 to 15 (2019 to 2023) provided the opportunity to estimate nutrient and food intakes by calculating 'usual intakes' which is the accepted method to estimate population habitual nutrient and food intakes. Dietary data for years 1 to 11 (2008 to 2018) were collected over 4 consecutive diary days and so nutrient and food intakes were estimated by calculating the 'day average'.

When calculating 'day average' nutrient and food intakes the variance of the usual group intake is inflated by day-to-day variation in individual intake, resulting in misleading estimates of the prevalence of low or high intakes. With the collection of repeated 24-hour recalls in years 12 to 15, it is possible to eliminate the intra-individual variability of the data and thereby to obtain an estimate of the population usual intake distribution (Souverein and others, 2011). Several statistical procedures for estimating the usual intake distribution from repeated 24 hour recalls are available to enable the estimation of 'habitual' intakes. This enables more appropriate estimation of 'percentiles' or 'proportions above or below a threshold' compared with the 'day average' method. See National Cancer Institute Diet Assessment Primer and National Cancer Institute: Usual Dietary Intakes, the NCI method for more information.

Usual intakes of frequently consumed foods such as meat, fruit and vegetables and sources of protein, can be estimated well with short-term measures (2 or more days of a 24-hour recall). Short-term measures for infrequently consumed foods such as fruit juice, fish and sugar-sweetened soft drinks can result in zero intake being reported for many participants. Including an additional long-term measure such as a Food Frequency Questionnaire (FFQ) can help to more accurately capture usual intake for such foods. For example, with a dietary assessment protocol of collecting up to 4 recalls, a participant may record zero intake for a particular food on all recording days but may record in an FFQ that they consume the same food on average once a week. FFQ information has therefore been collected in Computer Assisted Personal Interview (CAPI) for a limited number of infrequently consumed foods (fish, white meat, fruit juice and sugar-sweetened soft drinks) and will be factored in alongside the short-term measure to improve the estimation of 'usual intake'.

The Multiple Source Method (MSM, Germany) is a web-based application which uses R code to perform the statistical analysis and was used to estimate 'usual intakes' of nutrients and foods for years 12 to 15. All valid recalls from all participants, regardless of the number of recalls obtained, were included in the analysis. The inclusion of participants with only 1 or 2 dietary recalls was made possible by the usual intake method 'borrowing' day-to-day variation estimates from other similarly aged participants with 3 or 4 recalls. This provides an advantage over the previous diary method which required at least 3 diet collection days for a participant to be included.

For some foods there were insufficient participants with 2 or more consumption days and so it was not possible to estimate the day-to-day variation required in the calculation of 'usual intakes'. The threshold set for this was 15% or more participants with 2 or more consumption days. For foods which did not meet this threshold intakes were estimated by calculating the 'day average'.

U.4 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[1] 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 was used to confirm the choice of descriptive statistic for each variable before it is used in this report.

U.4.1 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 table U.1 along with the reported descriptive statistics:

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

Table U.1 Descriptive statistics reported for variables that deviate from a symmetric and 'bell-shaped' distribution

Analyte	Distribution	Descriptive statistics	Descriptive statistics
Blood: Red cell blood folate	Positively skewed	Geometric mean	2.5th and 97.5th percentiles
Blood: Serum folate	Positively skewed	Geometric mean	2.5th and 97.5th percentiles
Blood: Unmetabolised (free) folic acid	High proportion of values below the limit of quantitation	Median	2.5th and 97.5th percentiles
Urine: Iodine concentration	Positively skewed but following WHO guidance on descriptive statistics (Institutional Repository for Sharing (https://iris.who.int/))	Median	20th and 80th percentiles

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.

- Foods: Sugar-sweetened soft drinks, Sugar confectionery, Chocolate confectionery, Sweet or savoury biscuits and cereal bars, Buns, cakes and pastries and Crisps and other savoury snacks.
- Nutrients: Vitamin A, Vitamin D.
- Blood analytes: Ferritin, Vitamin B6 (PLP), Vitamin B12.

U.5 Trends over time

This section outlines the statistical methods used to estimate the 'average change per year' in each outcome for urinary iodine and self-reported physical activity energy expenditure from years 1 to 15 of NDNS. Trends over time for foods and nutrients and blood analytes were not estimated due to the dietary assessment method change to Intake24 occurring from year 12 and the blood sample postal model change from year 13. These time trends are investigated in the Stage 3 evaluation report of the dietary method change and the blood sample transport report. The same weights and design variables as those used in the years 1 to 4 (combined), years 5 and 6 (combined), years 7 and 8 (combined) and years 9, 10 and 11 (combined) reports (with additional weights and design variables for years 12, 13, 14 and 15 (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 (4:2:2:3:4) to give a standardised sample size per survey year.[2]

[2] Although the weights were not specifically designed for this type of sub-group analysis, it was possible to use the years 1 to 15 weights and design variables for just 2 to 4 years' data (years 1 and 2, years 3 and 4, years 5 and 6, years 7 and 8, years 9, 10 and 11 or years 12, 13, 14 and 15), 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 6 survey periods

However, to use subsets of any other combination of years of the dataset, the weights and design variables would

The ‘average change per year’ were estimated through linear regression models across the age groups, overall and by sex (for all but the 1.5 to 3 years age group). Participants were grouped into quarters of a calendar year according to when their urine sample physical activity questionnaire data was collected, and this time variable 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 the ‘diet method change’ 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.5.1 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.5.2 Estimation of the ‘average change per year’

Linear regression models were used and the regression coefficients (which estimate the intercept and slope parameters for each age and sex group) use probability weighted least squares (Holt and others, 1980) and their covariance matrix was estimated using a Taylor linearization method (Binder, 1983). The slope parameter (along with the associated 95% confidence interval) estimates the ‘average change per year’ for each variable.

U.5.3 Diagnostic procedures

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

U.6 Socio-economic regression analysis

This section outlines the statistical methods used to estimate the average change per £10,000 of equivalised income in each outcome of key foods, nutrients and blood and urine analytes from NDNS years 12 to 15 combined.

The ‘average change per £10,000 of equivalised income’ 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 6 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 to 74 years and 75 years and over.

The statistical analyses were undertaken using the following 3 stages: exploratory analyses,

---

have to be reviewed to ensure that the subset of data is still representative of the UK population when the years 1 to 15 weights and design variables have been applied.

estimation of ‘changes per £10,000 of equivalised income’ 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.6.1 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.6.2 Estimation of the ‘average change per £10,000 of equivalised income’ for continuous variables

Linear regression models were used for continuous measurements of foods, nutrients and blood and urine analytes. The regression coefficients (which estimate the intercept and slope parameters for each age and sex group) use probability weighted least squares (Holt and others, 1980) and their covariance matrix was estimated using a Taylor linearization method (Binder, 1983). The slope parameter (along with the associated 95% confidence interval) estimates the ‘average change per £10,000 of equivalised income’ for each variable.

U.6.3 Estimation of the ‘average change per £10,000 of equivalised income’ 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 (Holt and others, 1980) and their covariance matrix was estimated using a Taylor linearization method (Binder, 1983). The slope parameter (along with the associated 95% confidence interval) estimates the ‘average change per £10,000 of equivalised income’ for each variable.

U.6.4 Diagnostic procedures

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

U.7 Socio-economic quintile summary for England

To assess the relationship between each outcome of key foods, nutrients and blood and urine analytes from years 12 to 15 combined of the NDNS with the Index of Multiple Deprivation for England (EIMD) descriptive statistics were calculated (using the same methods as described in section U.3 above) for each quintile of EIMD. These descriptive statistics were then compared with descriptive statistics calculated for each quintile of equivalised income for England participants.

U.8 General

The statistical analyses described above were performed using the survey package in the statistical program R (Lumley, 2012) (Lumley, 2004).

The statistical analyses described in this appendix are for descriptive purposes rather than analytical, that is, they are not intended to estimate the associations among many variables. Therefore, corrections for multiple comparisons were not necessary (or practical since thousands of statistical tests have been performed). 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) (Korn and Graubard, 1990).

Unless stated otherwise, only trends and differences found to be statistically significant at the five per cent level are identified as `significant`. In other words, differences as large as these have no more than a five per cent probability of occurring by chance. The term ‘significant’ is not intended to imply substantive importance.

References

Binder D A. `On the Variances of Asymptotically Normal Estimators from Complex Surveys`. International Statistical Review 1983: volume 51, pages 279–292

Holt D, Smith TMF and Winter PD. `Regression analysis of data from complex surveys`. Journal of the Royal Statistical Society A 1980: volume 143, pages 474 –487

Korn EL and Graubard BI. `Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni t statistics`. The American Statistician (1990): volume 44, pages 270 – 276

Lumley T. `Survey: analysis of complex survey samples`. R package 2012, version 3.28-2.

Lumley T. `Analysis of complex survey samples`. Journal of Statistical Software (2004): volume 9, issue 1, pages 1-19

Souverein O, Dekkers A, Geelen A and others. Comparing four methods to estimate usual intake distributions. European Journal of Clinical Nutrition 2011: Volume 65, Supplement 1), pages S92–S101.