Appendix U - Wales
Statistical methods for the trends over time (Years 1 to 9) and socioeconomic (Years 5 to 9) regression analysis for the NDNS Rolling Programme (RP)

U.1 Introduction

This appendix provides an outline description of the statistical methods used for the assessment of trends over time from Years 1 to 9 (2008/09 – 2016/17) and trends in relation to socioeconomic status of the participant's household (using equivalised income) for Years 5 to 9 (2012/13 – 2016/17) of the NDNS RP including estimating the difference of these trends between overlapping subpopulations (defined by country).

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 the 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. A detailed description of the weighting and sampling procedures is provided in appendix B.

U.2 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 9 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 Year 9) 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 (4:2:2:1) to give a standardised sample size per survey year.[1]

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 the standard NDNS 5 age groups, overall and by sex (for the age groups 4 years and over only). The age groups were 1.5 to 3 years (sex-combined only), 4 to 10 years, 11 to 18 years, 19 to 64 years and 65 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.

U.2.1 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.2.2 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.2.3 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[4] 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.2.4 Diagnostic procedures

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

U.2.5 Calculation of 9-year average change

Calculation of the 9-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 9 to get the average change over 9 years e.g.:
  - Average change per year = -0.2mg/day (a reduction of 0.2mg/day per year).
    Average change over 9 years is 9*-0.2 = -1.8mg/day
  - Average change per year = +0.2mg/day (an increase of 0.2mg/day per year).
    Average change over 9 years is 9*0.2 = +1.8mg/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 9 times (i.e. calculate it to the power of 9) and then convert back to a percent change over the 9 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 9 years is ((0.97)^9-1)*100 = -24%
  - 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 9 years is ((1.03)^9-1)*100 = +30%

Average 9-year changes for each outcome of key foods, nutrients and blood analytes are provided in Excel tables U.1-U.4.

U.3 Socio-economic 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, blood and urinary analytes from Years 5 to 9 of the NDNS RP. The same weights and design variables as those used in the Years 5 and 6 (combined) and Years 7 and 8 (combined)

reports (with additional weights and design variables for Year 9) 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 (2:2:1) to give a standardised sample size per survey year.[1]

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 the standard NDNS 5 age groups, overall and by sex (for the age groups 4 years and over only). The age groups were 1.5 to 3 years (sex-combined only), 4 to 10 years, 11 to 18 years, 19 to 64 years and 65 years and over.

The statistical analyses were undertaken using the following 3 stages: exploratory analyses, 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.3.1 Exploratory analyses

The observed distribution 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.3.2 Estimation of changes per £10,000 of equivalised income for continuous variables

Linear regression models were used for continuous measurements of food, nutrient, blood and urinary analytes. The regression coefficients (which estimate the intercept and slope parameters for each age/ sex group) use probability weighted least squares 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 £10,000 of equivalised income for each variable.

U.3.3 Estimation of changes 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 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 £10,000 of equivalised income for each variable.

U.3.4 Diagnostic procedures

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

U.4 Comparison of trends between overlapping subpopulations

The comparisons between the NDNS RP Wales trends and the NDNS RP trends for the UK as a whole involve comparing the slope parameter from the regression model between overlapping subpopulations. The mean difference is the subtraction of the slope parameter for Wales and the slope parameter for the UK. However, estimation of the standard error of the mean difference requires consideration of the overlapping of the sample.

For illustration, consider the comparison of slope parameter of total fruit consumption in grams between Wales and the UK across age groups, where Wales is a subset of the UK as a whole. Suppose the slope parameter for Wales and the UK are $\bar{y}_1$ and $\bar{y}_2$, respectively. The standard error of the mean difference $d = \bar{y}_1 - \bar{y}_2$ can be calculated using the formula below:

$$s.e.(d) = \frac{r}{t}\sqrt{var(\bar{y}_{1\notin 2}) + var(\bar{y}_1)}$$

Where r refers to the weighted sample size of the UK after excluding Wales and t refers to the weighted sample size of the UK as a whole; $var(\bar{y}_{1\notin 2})$ represents the variance of the slope parameter for the UK excluding Wales and $var(\bar{y}_1)$ represents the variance of the slope parameter for Wales.

The Z-score for testing whether the mean difference is significantly different from zero can be obtained by:

$$Z = \frac{d}{s.e.(d)}$$

U.5 General

The statistical analyses described above were performed using the survey package in the statistical program R.[5,6]

Plots showing the estimated regression line are provided along with individual participant responses and means and proportions per calendar year or income decile. These means and proportions are just for illustration and are not intended to be used as population group means and proportions as they are calculated using only a single calendar year or a single income decile so will be based on a small number of participants. Population group means, based on paired survey years, and so a larger number of participants, are provided for Years 1 to 8 in the Years 7 and 8 (combined) report.

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 PSUs (Primary Sampling Units).[7]

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.

[1] Although the weights were not specifically designed for this type of sub-group analysis, it was possible to use the Years 1 to 4, Years 5 and 6, Years 7 and 8, and Year 9 weights and design variables, as:

- The selection weights correct for any differences in sampling strategy across survey years,
- We did not find evidence that response behaviour had changed significantly across the 9 survey years.

Re-scaling of the weights ensures that each survey year contributes equally in the analysis. 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 9 weights and design variables have been applied.

[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] Skinner, C.J. (1989) Domain means, regression and multivariate analysis. In *Analysis of complex surveys* (eds C.J. Skinner, D. Holt and T.M.F. Smith). Chichester: Wiley.

[5] 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

[6] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[7] 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.