NCT04816955

DIETARY RECOMMENDATIONS FOR REDUCING FREE SUAGR INTAKES

DATA HANDLING AND ANALYSIS PLAN

Date completed: 28.06.23

DATA HANDLING

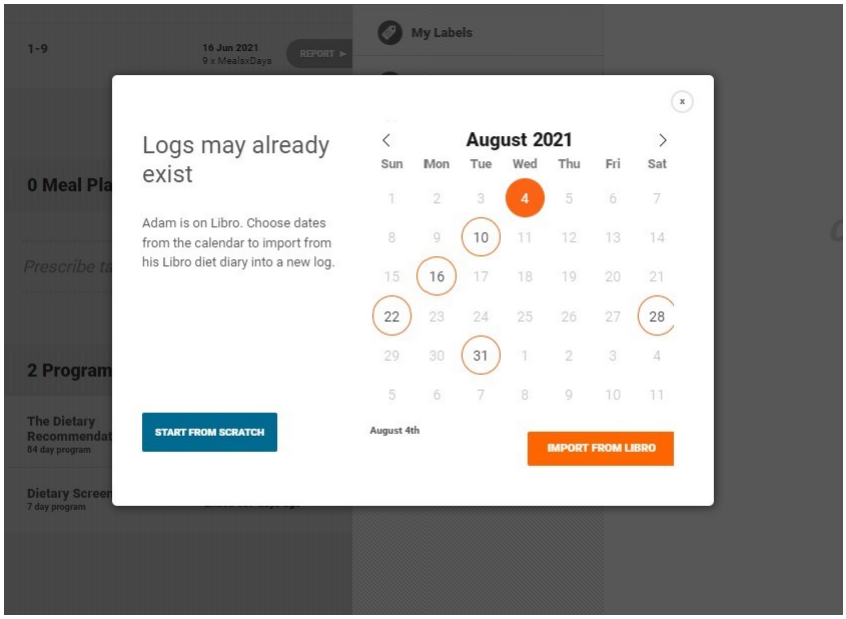
All data were collected in accordance with the study protocol. Missing data were monitored throughout the study and recovered at the time where possible. All collected data were cleaned and made ready for analysis only once the study was complete.

Dietary data

Dietary data were collected using diet diaries, completed via the Nutritics and associated Libro app platform. Eighteen daily diaries, plus three at baseline and three at Week 12 were requested from each participant.

Using these diaries, separate logs for each participant were created as follows.

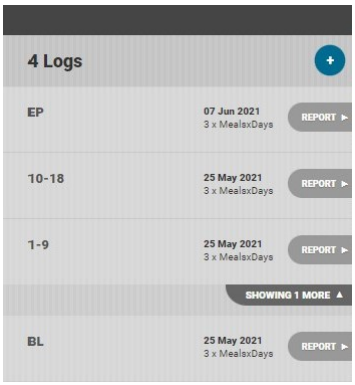
Diet diaries recorded on specific days as outlined in a master excel document were matched to specific dates of diet diaries as recorded by the participant. This is shown in the image below:



Dietary logs for analysis refer to the recording of multiple days of diet diaries generated into one report. For example, a 3 day log would use diet diaries on the 4th, 10th and 16th of August 2021 as shown in the image above.

Diet diaries were sorted and named as the following for all participants before analysis:

- BL – baseline data from the 3 days of dietary screening
- 1 to 9 – Diet diaries 1-9 as recorded by the participant. If one or more days are missing there would only be 8 or less days of diaries selected. If one or more extra days were used to complete the diary, then there may be 10 or more diaries selected.
- 10 to 18 – As shown above but for diaries 10-18
- EP – endpoint data from the 3 days of the final week on the study.

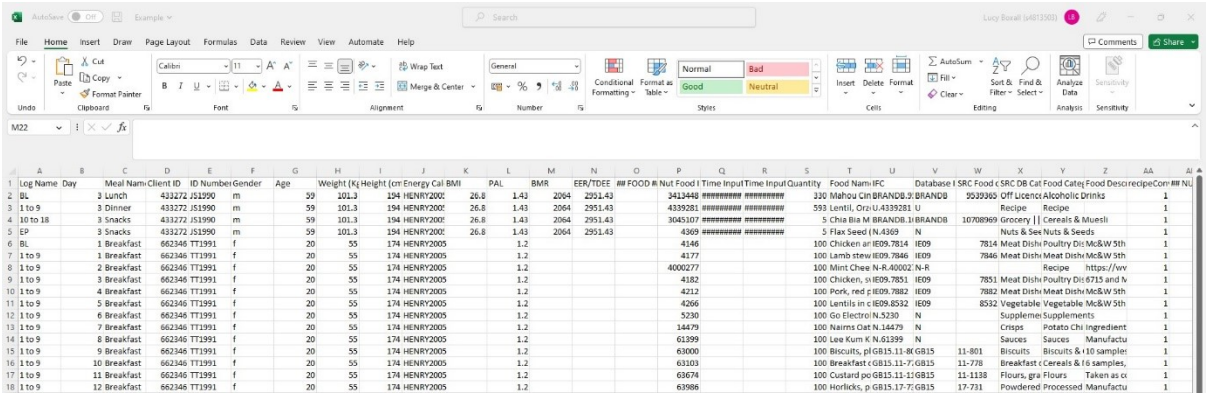


During the creation of dietary logs on Nutritics for analysis, when data for specific dates were missing or extra days were used – this was simultaneously recorded in a separate excel document as shown below.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	ID	Baseline	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Endpoint complete	Percentage complete	Percentage incomplete	Dropout /
2	AA1987	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	9	100.00%	0.00%	0
3	AB1965	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	9	100.00%	0.00%	0
4	AB1984	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%	100.00%	1
5	AB1989	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	9	100.00%	0.00%	0
6	AC2003	3	3	3	3	3	3	3	3	3	3	3	3	3	3	0	3	3	3	3	3	3	3	9	95.24%	4.76%	0
7	AD1972	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	9	100.00%	0.00%	0

Where data were missing, 0 was input. If data for a single dietary entry was recorded over multiple days, then a note was made on the participants row to adjust numbers down eg: day 9 and 10 show data for just day 9, day '10' in the Nutritics document will then be adjusted down to 9.

Once all logs for each participant had been created on Nutritics, they were batch exported into a CSV excel document. An example is shown below using hypothetical data.



Once exported log days are in the original numerical order for each log. A backup of this data is saved as file CSV1 with no edits. A new copy file called CSV2 is then corrected using the previous spreadsheet shown above which compares dates of diaries logged and whether that information has been input. Baseline diaries were corrected to days, 0.1,0.2,0.3. Diaries 1 to 9, corrected to 1,2,3,4,5,6,7,8,9 and 10 to 18 to 10,11,12,13,14,15,16,17,18. Endpoint diaries were input as 19,20,21.

A tertiary copy of CSV1 was made in order to create a list of individual food items logged, called CSV3. The 'remove duplicated' function was run on all food item codes in CSV3 resulting in 10294 uniquely coded foods items. A total of 7771 items did not have information recorded for free sugars. All nutritional information for all food items was updated from their original quantity in grams to 100g.

All food and drink items logged were checked individually using a variety of filters to analyse if the product was likely to contain free sugars. Filters included, 'text search' (broccoli), 'food name', 'food category', 'energy', 'kcal', 'carbohydrate' and 'sugar content'. Often multiple filters were used to increase the efficiency of recording, eg: search 'broccoli' and filter for lowest energy in kcal. First, it was decided which foods needed to be recorded, for example 'Sainsburys frozen broccoli' was unlikely to contain free sugars and therefore the original food code was kept with a 0 put into the free sugars column. For foods that did contain free sugars such as 'Cadbury mini eggs' these were recoded to another substitute food that had the full information on free sugars. The substitute coded item was selected by comparing the average of that food/drink item, and the comparative

nutritional values of energy (kcal), carbohydrates and sugars. For specific foods and brands, where it was unknown if the item was likely to contain free sugars, the ingredient list was researched online, using online grocery stores such as Ocado, Sainsburys, Tesco's, Asda and occasionally the brands own website. Every effort was put into retaining as many of the original food items as previously logged.

Once all items without free sugar data had been analysed to see if recoding was needed another spreadsheet was created, and these data were input into a separate sheet within CSV2. All 67406 individual food items were updated to a 100g value using the same method as previously mentioned. Then the 'XLOOKUP' function was utilised to identify which codes needed to be replaced and those that remained the same. The 'delta' function in a helper column was then used to compare the two columns of 'New' and 'Old' food code. All original items to be kept were then filtered and hidden in the spreadsheet. Only items which required recoding were now showing in the spreadsheet. The 'XLOOKUP' function was again used to match codes and replace all food names and nutritional values for recoded foods. Once completed, the spreadsheet was copied and pasted to remove all formulas and just retain the values. All substituted and non-substituted food items were then updated to their original quantities. The CSV2 file was then saved for backup.

This updated data from CSV2 was copied into a new file named CSV4. In CSV4 a helper column combining the participants ID number and log day (eg: JS19900.1, combination of JS1990 and 0.1) was used in combination with a new sheet to summarise all nutrients for each participant and each diet diary day. This resulted in 5120 unique daily totals across all nutrients. The CSV4 file was then saved as complete. Including all participants and screened non-participants a total of 70747 individual diary entries on the study via Nutritics Libro. Of these data, 67546 were participant entries and 3201 from non-participant screening diaries, with a total of 10757 unique food items logged.

Sweetener and sweet food counts

Foods that contained sweeteners, were high, medium or low in sugar were separated into counts for each diet diary. These counts were then added into 5 separate scores for the summation of foods consumed in each category for baseline, timepoints 1-4 and endpoint.

ALL DATA - MULTIPLE IMPUTATION

Main analyses were completed on an intention to treat basis as the gold standard for representing the applicability of the intervention in a real world setting rather than an ideal outcome. A variety of techniques to replace/analyse RCT results with missing data have been previously utilised including complete case analysis (CCA), single imputation (SI)(by last observation carried forward, mean or regression), multiple imputation or mixed-effect models (Bell et al., 2014; Li & Stuart, 2019; Ren et al., 2022). CCA and SI methods were discounted due to CCA not being reflective of real-world data, reducing statistical power and SI lacking uncertainty and most likely underestimating variance.

Instructions for the steps used to complete MI and the imputation sequence are included below. As multiple imputation is completed in the order of the variables within the analysis, the variables list was set as randomisation variables first, then nutrient values, anthropometrics and finally questionnaire data in the order that they were completed. All baseline variables were used as predictors, all endpoint variables used as imputations only due to the monotonic pattern.

Current literature suggests the optimal number of imputations varies between 5-20, with limited benefit over 5. Additional literature supports the number of imputations matching the percentage of attrition. Therefore, the number of imputations for the main analyses was set at 20 representing both the upper recommendation and percentage attrition.

Updates to SPSS now allow random selection of the closest of over >1 donor. Existing methodology and the amount of imputations suggest imputations via Predictive Mean Modelling are selected from the 5 closest donors making this more statistically robust than previous methods.

Before completing multiple imputation, the pattern of missing data was checked. Endpoint data presented as missing in a monotonic pattern. When imputing missing data values in SPSS the custom 'fully conditional specification (FCS)' in methods was selected over monotone. This was because FCS is suitable for both monotone and non-monotone patterned data and introduces Bayesian stochastic regression imputation as part of the predictive means multiple imputation methods. The benefit of this method results in imputed data being taken from both a donor and then adjusted to introduce a level of variability rather than just a copied result. Maximum iterations were set as 50 with FCS convergence checked post imputation. For all applicable statistical tests, analyses were run on each imputed dataset, with values pooled post analysis. This is because a pooled dataset prior to analysis does not just represent the average but retains the variability of the individual imputed datasets.

Where SPSS did not automatically pool results, Rubin's Rules for combining single estimates were used. The same multiple imputation guidelines were used for the basis for the imputation of all missing data. Where variables presented as having missing data, they were set as 'impute only' to ensure all predictor variables were not based on imputed data. The only difference was in one baseline waist circumference value that was missing in the original data and used in the 2nd MI predictors of the subsequent variables. This study took the premise of not generating imputations based on imputed data values. The same multiple imputation model was run for the initial analyses, adherence and variables completed by only lab participants. Multiple imputation for the adherence data used the same methods and predictors as for the main data set. The step by step methods used are included below. Adherence data points of each free sugar intake at each dietary data point were not included as predictors due to the fact that at least 1 time point needed to be imputed at each variable time point.

Instructions for multiple imputation in SPSS

1. Go to transform – random number generator. Select and set the active generator initialization at a fixed value of 950.
    a. (This step is to provide a starting point/seed for the MI to enable results to be reproduced)
2. Then Analyze – multiple imputation – impute missing data values.
    a. Variables tab
        i. Input all variables for the model in the order they are to be used (this can be found in the output document and also to summarise at the end of this doc) but is all endpoint and corresponding baseline variables, therefore excluding demographics. But including gender. (The order is central for reproducing)
        ii. Create a new data set – name ComA1_Impu
        iii. Change imputations to 20
    b. Method tab
        i. Custom – select fully conditional specification (set to 50 iterations)
        ii. Model type – change to PMM – chose from 5 closest (as is default)
    c. Constraints tab
        i. Scan data
        ii. Under define constraints – set all Endpoint variables as impute only.
            1. Instead of clicking use tab and I, plus directional keys to change this.
    d. Output tab
        i. Tick imputation model, descriptive statistics and create iteration history – name ComA1_iter
    e. Double check – click ok
3. Save the imputed and iterations files.

Sequence for multiple imputation.

BLNFS,BLDGendermale1female2,BLBMI,BLNEnergykcal,BLNCarbohydrateg,BLNProteing,BLNFatg,BLNSugarsg,BLNFreeSugarsg,BLNFibreg,BLNSaturatedFatg,BLSodiumNamg,BLWeightkg,BLWaistcircumferencecm,BLLTEShealthscore,BLFCQ1Health,BLFCQ2Mood,BLFCQ3Convenience,BLFCQ4SensoryAppeal,BLFCQ5NaturalContent,BLFC1Q6Price,BLFCQ7WeightControl,BLFCQ8Familiarity,BLFCQ19EthicalConcern,BLSQPC1Personalimpact,BLSQPC2Personalmanagement,BLSQPC3Nonchalance,BLSQPC4Negativity,BLSQPC5Perceivedunderstanding,BLSQPC6PerceivedNonautonomy,BLSF36PCS,BLSF36MCS,BLTFEQCR,BLTFEQUE,BLTFEQEE. **Followed by imputation only variables*

Get pooled descriptives for analysis / comparison to my values.

1. Data - Split file
    a. Click split file in ComA1_imp
    b. Click compare groups
    c. By imputation number_variable
    d. Ok
2. Analyze – Descriptive statistics
    a. Descriptives
    b. Input all BL and EP variables apart from BL OG FS%.
    c. Scroll to bottom of output table for pooled means – SD will be calculated from average of 20 imputations at a later date and are not shown here.

BLINDING

The main researcher had no knowledge of allocated groups of participants and was blinded during all data handling and imputation.

DATA ANALYSIS

Outcomes

Primary outcomes were % TEI from free sugars (%FS), and adherence to the recommendations at 12 weeks. Secondary outcomes were dietary profiles (daily energy intake, dietary composition, food consumption), anthropometry, sweet food perceptions and preferences, sweet food choice, attitudes to sweet foods, attitudes towards eating behaviour, motives for food choice, knowledge and lifestyle variables, quality of life and adverse events at week 12; change between baseline and week 12 in key outcomes (%FS, daily energy intake, sugar-sweetened and sweetener-sweetened food consumption, anthropometry), %FS, adherence, dietary profiles and barriers and facilitators towards intervention adherence and success/failure in achieving the recommendations (subset only) at weeks 1, 2, 4 and 8. Demographic variables, and sweet taste sensitivity, sweet liker status and PROP (bitter taste) status were also assessed at baseline, to aid in the interpretation of all outcomes, considering associations between these variables and sweet taste perceptions, liking and food intakes (Jayasinghe et al., 2017; Yeomans et al., 2009).

Three distinct analyses were pre-specified in advance: 1) Analyses of the population as a whole to investigate the effects of the three different types of dietary recommendation versus control; 2) Analyses of the effects of the dietary recommendations in different population subgroups, and 3) Investigation of the barriers and facilitators to success.

Data presentation

Data will be presented as mean ± standard deviation in all cases, unless stated. Values will be considered significant at the $p<0.05$ level. Before completing analyses, data will be checked for normality. The data set is large, N=242, thus a normal distribution will be assumed. Where assumptions of homogeneity of variance are violated (Levene's test $p<0.05$), statistics will be taken from analyses conducted assuming non-equal variances. Bonferroni corrections for multiple testing will be applied to data where tests were not pre-specified in the study protocol. Where multiple regression analyses are used, all variables will be checked for multicollinearity using correlation cut-offs of <0.7. Pearson correlation values will be used for all continuous variables, and Spearman correlations for categorical variables.

Participant characteristics

Details of study completion will be provided and a CONSORT diagram created detailing number randomized to each group, number completing the study in each group, drop out and number included in all analyses.

Characteristics will be given for the whole sample and for each trial group.

Analyses One: Quantitative Data

For analyses one, multiple regression analyses will be used. Analyses will be conducted for both primary outcomes, for key secondary outcomes - daily energy intake, sugar-sweetened food intake, LCS-sweetened food intake, body weight, waist circumference, and adverse events; and change (baseline - week 12) data for these outcomes; and for other secondary outcomes (dietary

composition, sweet food perceptions and preferences, sweet food choice, attitudes to sweet foods, attitudes towards eating behaviour, motives for food choice, knowledge and lifestyle variables, quality of life) if differences are found in the sample as a whole from baseline to week 12. This will reduce the number of exploratory tests undertaken. Differences from baseline to week 12 will be investigated by paired t-tests. Primary outcomes a d outcomes associated with dietary composition will also be investigated in the same manner at weeks 1, 2, 4 and 8.

Primary models will include gender, age, control/intervention group, baseline %FS, baseline body weight (as randomization variables), and baseline variable of interest. In analyses for %FS, sugar-sweetened food intake, LCS-sweetened food intake, body weight and waist circumference, TEI at Week 12 and physical activity at Week 12 will also be included. Secondary models will also include any additional variable that correlates with each outcome when assessed independently. All variables measured will be considered as secondary variables. Effects for all continuous variables will be investigated using linear multiple regression. Effects for adherence will be investigated using logistic regression to predict adherent vs not adherent.

Analyses Two: Quantitative Data

For analyses two, analyses one will be repeated for demographic groups where significant associations for these variables were found in analyses one. Analyses will only be conducted if significant associations are reported. If no significant associations with demographic variables are found, these analyses will not be completed. These analyses are intended to demonstrate differing effects of the intervention/s in different population groups.

Analyses Three: Qualitative Data

Analyses three will investigate the qualitative data collected from a subset of participants using solo interviews. These interviews were scheduled for completion in a sample of 80 participants from different intervention groups and at different time points. These analyses will be undertaken using a thematic framework analysis. First, a random sample of interviews will be analysed inductively using thematic analysis to identify codes. These codes will be used to create a framework and the framework then applied to all interviews.

Framework analysis has been chosen as a suitable analytical method for evaluating the qualitative interview data gathered, due to an interest in individual experiences and seeks to identify the different factors (barriers or facilitators) that contributed to successful or unsuccessful dietary change (Braun & Clark, 2013). The practical methods for conducting framework analysis follow a methodology as outlined below.

Framework analysis (Braun & Clarke, 2006; 2013; Gale et al, 2013).

Phase 1 'Familiarizing yourself: Transcribing data (if necessary), reading and re-reading the data, noting down with your data: initial ideas'.

- Audio data is transcribed, using automated timestamps and a published notation system in NVIVO software.
- To ensure high quality audio and transcription practises, Poland's (2001) '*Strategies for Ensuring High-Quality Tape Recording'* will be used.

- All transcription is completed orthographically to produce a verbatim account. Data will be re-read and checked to ensure accurate transcription. Once the researcher has completed the transcription, data will be anonymised, and the audio file destroyed.

Phase 2 'Generating initial codes: Coding interesting features of the data in a systematic fashion across the entire data set, collating data relevant to each code.'

- Using a process of complete coding, initial codes will be generated after the transcription of the first 12 interviews.
- The data set will only be considered saturated, once fewer than 5% of new codes are identified in on-going analyses. Two additional transcripts will be required at this point to confirm saturation.
- Once the data saturation is confirmed, inter-rater reliability will be tested.
    - Two randomly (computer generated number selection) selected transcripts and the coding book will be provided to a second coder for this test.
    - Only once substantial coding agreement between the researcher and second coder has been achieved (κ = >0.61)(80) can phase 3 begin.

Phase 3 'Developing a working analytical framework.'

- After reaching saturation as agreed above, the researchers will meet and agree the codes for all subsequent transcripts. These codes are then grouped together into categories or themes.

Phase 4 'Applying the analytical framework'

- All transcripts are to be coded according to the agreement in phase 3.
- All candidate themes will be reviewed and checked to ensure a coherent theme pattern is present.
- Then the whole data set will be reviewed providing answers to these questions.
    - Do the themes work in relation to the whole data set?
    - Does any re-coding of data within themes need to be completed?
    - Has anything been missed from the initial coding?

Phase 5 'Charting data into the framework matrix'

- Data from transcripts will be charted into an excel spreadsheet outlining each interview and the extracted quotes (and location reference) for each code.

Phase 6 'Interpreting the data'

- Characteristics and difference between groups and interview timepoints will be analysed.
- Mapping connections between relationships and categories.
- A 15-point check list from Braun & Clarke 2013 and consolidated criteria for reporting qualitative research (COREQ) will be used in the write-up and reporting of this qualitative research.