# **MESA Exam 1 Diet Correction Documentation**

Contact: Craig Johnson ([wcraigj@uw.edu](mailto:wcraigj@uw.edu))

Date: 05/10/2015

## **1. Conditions of data use (in addition to MESA data access policy):**

1. Coordinating Center will ensure that MESA Investigators who request access to MESA Exam 1 Diet data have initiated **discussion of their planned use of the data with Members of the MESA Diet Working Group** prior to gaining access to the data. Members of the Diet working group with specific knowledge of the situation include Alexis Frazier-Wood, David Jacobs, Lyn Steffen, Majken Jensen, Marcia Otto, Robyn McClelland, Rebekah Young, Bridget Kruszka, and Craig Johnson.
2. **Coordinating Center tracks specific access and use of the data** so that new information concerning its use and analysis can be easily disseminated to users.
3. While investigations of the corrected data have not identified major concerns in previously published results to date, it is important that **users of these data remain vigilant and report concerns and recommendations to the Coordinating Center at** [**wcraigj@uw.edu**](mailto:wcraigj@uw.edu) (potentially for disseminating as noted in condition 2 above).

## **2. Background**

September 2014 and February 2015 reports to MESA Steering Committee provide detailed description of a coding error that was apparently systematically applied to a subset of Exam 1 food frequency variables and for a subset of the patients. The following two linked presentation documents provide a more detailed description of the issue.

<http://www.mesa-nhlbi.org/MesaInternal/PresFeb2015/18_Johnson_MESAe1FFQ_02252015.ppt>

<http://www.mesa-nhlbi.org/MesaInternal/PresSep2014/Johnson_MESAe1FFQ_09172014.ppt>

In short, 19 Food frequency variables, 14 Food group variables, and likely all nutrient variables were noted to have at least some underreporting in a subset participants (that depended on participant response to modifier questions).

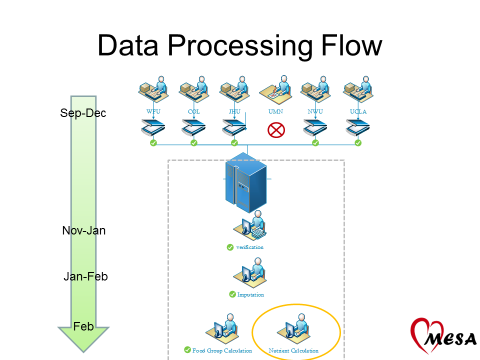
|  |  |  |
| --- | --- | --- |
| **Food Frequency variable** |  | **Food Group variable** |
| Apples |  | Fruit |
| Orange Juice |  | Fruit Juice |
| Other Juice |  | Legumes |
| Lettuce |  | Other vegetables |
| Spinach |  | Leafy green vegetables |
| Potato |  | Potato |
| Dark Bread |  | Whole grains |
| Crackers |  | Salty snacks |
| Chips |  | Yogurt |
| Cheddar cheese |  | High fat dairy |
| Yogurt (flavored) |  | Cottage Cheese |
| Yogurt (plain) |  | Red Meat |
| Cottage cheese |  | Processed meat |
| Hamburger |  | Desserts |
| Ham |  |  |
| Chile |  |  |
| Sausage |  |  |
| Chocolate donuts (cake) |  |  |
| White donuts (cake) |  |  |
|  |  |  |

## **3. Data Correction**

Original data processing required that the original Food Frequency Questionnaire paper forms be sent to the Diet Analysis Center for data processing. Upon completion of the dataset, paper forms were unfortunately destroyed which made verification of the original responses more difficult. Four of the six Field Centers saved photocopies of virtually all completed questionnaires and saved these copies in the participant charts. One center saved photocopies on approximately 25% of their participants and one center did not save a photocopy for any of their participants.

The five Field Centers who maintained photocopies of the questionnaire pulled those copies from their chart archives for scanning and submission to Coordinating Center for further processing. Coordinating Center set up a semi-automated form reading application with 100% verification of the affected fields. Food frequency, food group, and nutrient variables were subsequently recoded for participants where a form was available for reprocessing. Fields that were identified not to be affected by the error were not changed during this process.

In any instance where a form could not be located, single imputation methods are used as described later in this document.

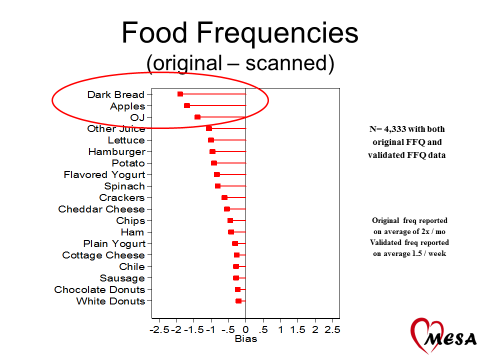


**Data Correction Summary**

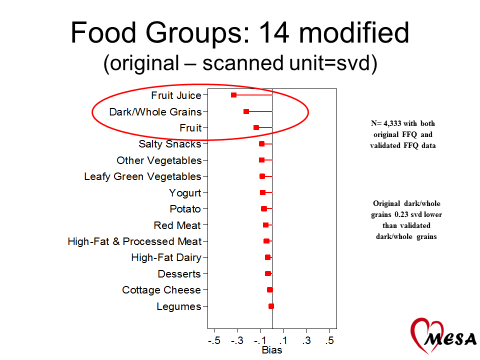
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Original | Corrected | Scans | Imputed |
| 3:WFU | 973 | 1,022 | 998 | 24 |
| 4: COL | 909 | 1,049 | 1,038 | 11 |
| 5: JHU | 955 | 973 | 239 | 734 |
| 6: UMN | 1,006 | 1,006 | 0 | 1,006 |
| 7: NWU | 1,119 | 1,164 | 1,053 | 111 |
| 8: UCLA | 1,274 | 1,317 | 1,299 | 18 |
| Total | 6,236 | 6,531 | 4,627 | 1,904 |

## **4. Comparison of scanned versus original data (excludes imputation)**

As predicted, frequency of modified foods (those affected by the error) consumed were lower in the originally released dataset. The most impacted food was dark bread with apples and orange juice also being similarly affected. Using dark bread as a more specific example, the original data had an average frequency reported of approximately twice a month while the corrected (rescanned) data had an average frequency of 1.5 times a week.



As predicted, servings per day of modified food groups (those including frequencies affected by the error) consumed were lower in the originally released dataset. The most impacted food group was fruit juice with Dark/Whole grains and fruit also being affected. Using fruit juice food group as a more specific example, the original data had an average serving per day that was 0.23 servings lower than the corrected (rescanned) data.



## **5. IMPUTATION**

**Methods**

All FFQ and serving size values were imputed using sequential chained regression implemented in Stata 12 using the MI program. This method allowed the imputed values to match the metric of the observed data, which was necessary for the food group and nutrient calculations. Each of the 19 FFQ items and corresponding serving size items were imputed in the same model.

The FFQ items were imputed using multinomial regression or ordinal regression (when appropriate). The serving size items were always imputed using multinomial regression. The imputation model for the 19 sets of items varied from question to question. In all cases, the imputation model accounted for basic demographics. Other auxiliary variables were selected on the basis of having a high correlation with the FFQ or serving size items. Some variables required separate imputation models to be run by race/ethnicity.

**Evaluation**

To evaluate whether imputing site-specific missingness appeared to be a valid approach, we created an artificial situation (“Scenario 1”) that mimicked the observed pattern of missing data. Starting with the validated data (n=4,229), we added 30% hypothetical site missingness. All participants from WFU were treated as missing, as well as non-Black participants from COL. This mimicked the actual missing data quite well based on the overall and race-specific proportions of missingess. Data were imputed and complete case and imputed estimates were compared to the “true” (known) distributions. All 19 FFQ and serving size items were tested. The distribution of the imputed FFQ items and calculations of servings per day were found to be considerably less biased than complete case analysis. “Scenario 1” was also applied to multivariable replication analysis for two diet data papers. Results consistently showed that the imputed data were superior to complete case analysis and showed small amounts of bias.

To evaluate how the actual pattern of missingness might influence the imputation approach, we created a second artificial situation (“Scenario 2”) that applied the observed pattern of missing data to FFQ items that were not subject to the modifier question error. 10 items were selected based on nearby proximity to one of the 19 FFQ items on the paper questionnaire. The imputation model from the corresponding FFQ item was used to impute each of the 10 items. As with the first scenario, complete case and imputed estimates were compared to the true distributions. The imputed values showed less bias compared to complete case analysis, and low levels of bias overall.

## **6. FUTURE WORK & RECOMMENDATIONS**

The next step in evaluating the imputed values is to compare a single imputed dataset to multiple imputed datasets.

The Coordinating Center will continue to validate and re-analyze published diet data papers.

When the imputed diet data become available for analysis, the Coordinating Center recommends that authors consider including a member of the MESA Diet and Nutrition Working Group as a co-author.

The MESA Diet and Nutrition Working Group will work to publish a paper detailing the methods used for imputation and results of the validated papers.