Digital Health Innovation for Consumers, Clinicians, Connectivity and Community
A. Georgiou et al. (Eds.)

doi:10.3233/978-1-61499-666-8-34

Development of an At-Risk Assessment Approach to Dietary Data Quality in a Food-Based Clinical Trial

Vivienne GUAN[a], Yasmine PROBST[a], Elizabeth NEALE[a], Allison MARTIN[a] and Linda TAPSELL[a]
[a] *School of Medicine, Faculty of Science, Medicine and Health, University of Wollongong, Australia*

Abstract. Accurate and valid dietary data is the basis to investigate diet-disease relationships. Potential data discrepancies may be introduced when collecting and analysing data, despite rigorous quality assurance protocols. The aim of this study was to identify at-risk areas of dietary data in a food-based clinical trial. Source data verification was performed on a 10% random sample (n=38) of paper-based baseline diet history interview records in a registered clinical trial. All items listed in the source data underwent 100% manual verification based on the food input data from FoodWorks nutrient analysis software. Food item discrepancies were explored using food categories and summarised based on meals. The differences in identified discrepancies for energy and macronutrient output generated from FoodWorks software between previously entered data and re-entered data were compared. An overall discrepancy rate of 4.88% was identified. It was found that dinner intake data were more prone to discrepancy incidences than breakfast, lunch and snacks. Furthermore, assessing intake based on reported quantity and frequency may be more effective to correct discrepancies for quality improvement. Therefore, the dinner meal appeared to be an at risk area of dietary data. The method implemented in this study offers a systematic approach to evaluating dietary data in a research setting.

Keywords. Source data verification, data quality, clinical data risk, diet history

Introduction

Diet plays a significant role in the development of many lifestyle-related diseases, such as cardiovascular disease, cancer, and type 2 diabetes mellitus [1]. Dietary data is used to describe food intake at the individual level [2] and high quality data is required to adequately reflect an individual's dietary intake to investigate diet-disease relationships. Dietary data is collected by applying validated dietary assessment methods, such as the diet history interview [2]. The collected data is often recorded on a paper-based case report form (CRF), and transcribed to a database for analysis supported by food composition tables.

Data entry errors are commonly found in clinical research databases [3]. Source data verification (SDV) is the procedure of ensuring that data accurately matches the original source data documents [4]. Although SDV has been reported to be time consuming and costly [5], it may provide critical evaluation of the processes related to data derivation workflow for dietary data quality improvement. Therefore, this study

applied SDV to explore dietary data entry discrepancies, with the aim to identify at-risk areas of dietary data entry within a lifestyle clinical trial.

1. Methods

1.1. Dietary Data Collection and Entry

Participant diet history interview records from a registered clinical trial were the basis for this work. Details of the clinical trial have been described elsewhere [6]. Dietary intake data reflecting usual weekly food consumption was collected by Accredited Practising Dietitians (APDs) during an open-ended interviewer-administrated interview. The meals, intake of food items, quantity and frequency were recorded on paper-based diet history interview CRFs (source data). All records were transcribed to FoodWorks Professional nutrient analysis software (Xyris, QLD, Australia, Version 7, 2007). Foods and quantities were transcribed by selecting items from drop-down lists in the software supported by the AUSNUT 2007 food composition database [7]. Where appropriate, new recipes of dishes and foods were created by dietitians and added to the database to accurately reflect participant reported intakes. Intake frequency was also transcribed to reflect the variations. The analysis automatically calculated intake frequency as an average intake per day. For example, consuming spaghetti bolognaise (1 cup as 1597kJ in FoodWorks) one time per week, automatically produces an average daily energy contribution of 228kJ (1597kJ/7=228kJ). In order to accurately estimate intake, total intake frequency of main meals were verified to equate to one on average meal per day.

1.2. Food-based Classification for Meals

Breakfast, lunch, dinner and snacks were used to group eating occasions (meals) during the SDV process. Other smaller meals, beverages and food frequencies were grouped together as snacks. Meal–based food consumption combinations (FCCs) were described as the sum of single food items consumed in the same meal or at the same time. For example, breakfast cereal and milk were often reported as being consumed at breakfast. The combination is counted as one breakfast FCC. Meal-based FCCs and frequencies for main meals were determined based on CRFs. All discrepancies were categorised according to food groups, based on a modified version of the 2011–13 Australian Health Survey food classification system at the major group level [8].

1.3. Discrepancy Classification and SDV Procedure

A 1% random sample (n=4) of CRFs from the same clinical trial was used to explore potential discrepancy types. Discrepancy types were established (Table 1).

A 10% random sample (n=38) of baseline CRFs from participants (n=377) were extracted. This method was based on the study by Mealer et al to investigate barriers to carry out large-scale randomised controlled trial [8]. The finding from the study conducted by Andersen et al also showed that selecting a random sample to conduct SDV assisted on the error reduction in a prospective clinical trial [9]. One researcher, an APD independent of data collection, performed the verification process to maintain

consistency. The data points in both the CRFs and the food output were summarised based on a single food item and values of its quantity and frequency. All items listed on the CRFs underwent a 100% manual verification check against the food output using the discrepancy types determined in the 1% sample. Discrepancies related to intakes of food items, the quantities, and frequencies were assessed using the food categories and summarised based on reported meals.

Table 1. Definitions and examples of discrepancy types

Discrepancy type	Definition	Example
Food items		
Incorrect	Recorded on CRF transcribed incorrectly or not related to food items to the database	Orange juice recorded on CRF but transcribed as orange to the database
Missed/missing	Recorded on CRF but not transcribed to the database	Recorded grated cheese 0.5 cup and not transcribed to database
Valid sourceless	Not recorded on CRFs though database contains an entry	Olive oil not recorded on CRF, database contains food item
Questionable	Mismatched between CRF and database or detail of ingredients for a dish are listed on CRF but pre-defined dish selected in the database	Recorded as bean stir fry in CRF, and transcribed as bean to database
Quantity		
Incorrect	Transcribed incorrectly	Recorded as one apple and transcribed as two apples
Valid sourceless	Not recorded on CRF though database contains an entry	Quantity of nuts not recorded on CRF, database record shows ¼ cup
Invalid sourceless	Total quantity of a number of food items recorded on CRFs but individual food quantities not recorded	Total amount of vegetable in beef stir fry recorded as 1 cup. Quantity of specific vegetables not recorded in CRFs and transcribed as broccoli ¼ cup, carrot ¼ cup, snow pea ¼ cup and onion ¼ cup.
Frequency		
Incorrect	Transcribed incorrectly	Recorded as once fortnight on CRF and transcribed as once per week

CRFs with identified discrepancies were re-entered into FoodWorks software. Those that could not be re-entered were kept as originally entered in the database. An inability to re-enter occurred if the discrepancies were of an invalid or valid 'sourceless' discrepancy type or if the intake of the specific food item, quantity and/or frequency were not recorded on CRFs (for example. if steak once per week was recorded on the CRF with no quantity, the entry could not be reentered due to the missing CRF quantity).

1.4. Statistical Analysis

Discrepancy rates were calculated based on the number of data points in CRFs. Invalid sourceless data for intake quantities were excluded from discrepancy analyses due to the total quantity of food items recorded on CRFs. CRFs that could not be re-entered were also excluded from statistical analyses. Statistical analyses were performed by using SPSS software package (Version 21, 2012, Chicago, IL). Normality of all data was checked using the Shapiro-Wilks test. A paired t-test for parametric data, and the Wilcoxon signed rank test for non-parametric data was used. Statistical significance was considered at $p<0.05$.

1.5. Results

Table 2. Relevant discrepancy type, number of discrepancies and discrepancy rate

	Sample (n=38)		**Discrepancy values re-entered (n=26)**		
	Number of discrepancies	**% discrepancies**	**Number of discrepancies**	**Number of discrepancy value re-entered**	**% discrepancy value re-enterable**
Food items					
Incorrect	18	0.6	16	10	56
Missing/missed	88	2.95	86	50	57
Valid sourceless	38	1.28	33	3	8
Questionable	31	1.04	29	5	16
Sub-total	175	5.87	164	67	38
Quantity*					
Incorrect	62	2.08	60	60	97
Valid soureless	100	3.36	72	0	0
Sub-total	162	5.44	132	60	37
Frequency					
Incorrect	99	3.32	99	99	100
Sub-total	99	3.32	99	99	100
Total	436	4.88	394	223	51

*Number of invalid sourceless of intake quantity was 232

A total of 8940 data points from 38 CRFs were verified. The total number of data points in the food output data was 8775, which was not significantly different from the data points on the CRFs (P=0.463).

A total of 436 discrepancies were identified, resulting in an overall discrepancy rate of 4.88%. The discrepancy rate of individual CRFs ranged from 0-60% (median 8%). There were 15 CRFs containing more than 10 discrepancies, and the discrepancies of 26 CRFs were able to be re-entered (Table 2).

The absolute differences in identified discrepancies for energy and macronutrient output between previously entered data and re-entered data are shown in Table 3. After re-entering discrepancies, the absolute differences in daily energy in three CRFs were found to be greater than 1MJ, thus, discrepancies of misreported data which were greater than 1MJ of energy intake was 8% (3/38). There was no significant difference between previously entered data and re-entered data for daily intake energy (p=0.123), protein (p=0.567), fat (p=0.058), carbohydrate (p=0.267) and fibre (p=0.188).

The greatest number of reported meal-based FCCs was for the dinner meal (median 6, range 1-11). The median number of breakfast and lunch FCCs were 3 (range 1-5) and 4 (range 1-7), respectively. A total of 16% (6/38) of accumulated total frequency instances of dinner were greater than eight (which should equate to seven ie. on average one time per week). Furthermore, a total of 48% (209/436) of discrepancies were identified for the dinner meal. Dinner had the highest discrepancy by meal for all discrepancy types.

Table 3. Absolute difference between previously entered data and re-entered data

		Daily intake (n=26)	**Breakfast intake (n=26)**	**Lunch intake (n=26)**	**Dinner intake (n=26)**	**Snacks intake (n=26)**
Energy(kJ/day)	Median(Range)	136 (3-3366)	0 (0-401)	34(0-3145)	113(0-660)	36(0-2954)
Protein (g/day)	Median(Range)	3 (0-45)	0 (0-3)	0 (0-47)	1 (0-14)	0 (0-41)
Fat (g/day)	Median(Range)	1 (0-56)	0 (0-7)	0 (0-63)	1 (0-10)	0 (0-32)
CHO (g/day)	Median(Range)	3 (0-69)	0 (0-10)	0 (0-4)	1 (0-10)	1 (0-63)
Fibre (g/day)	Median(Range)	0 (0-6)	0 (0-6)	0 (0-1)	0 (0-2)	0 (0-1)

2. Discussion

The use of SDV to investigate clinical trial data quality is not new, however no published studies have applied it to dietary data. The findings from this study contribute to the decision making process for at-risk areas which might be prone to discrepancies impacting on overall dietary data quality.

This study has demonstrated that the overall discrepancy rate of the dietary data set was 4.88%. Moreover, after re-entering discrepancies also identified 8% of these cases misreported greater than 1MJ of energy intake. As entering dietary data not only involves the numeric data entry, but also requires selecting the food items in the current available nutrient analysis software to accurately reflect the reported dietary intake. This process requires the high level of food knowledge and the high degree of professional judgement compared with other forms of data collection in a clinical trial. The discrepancy rate related to numeric data [10] and error reduction techniques by using different data entry methods, such as using number pad, cash register, modified number pad and number scrolled [11] may be unable to be employed by the dietary data. However, Clark et al demonstrated that discrepancy rates <10% are also

acceptable based on the verification of both numeric and descriptive data [12]. Therefore, our data set appears to be reliable for dietary analysis.

Data entry of the dinner meal may be prone to greater discrepancies. This may be due to its increased variety compared with other meals. FCCs increased, and homemade dishes were also more likely to be consumed at dinner. This may indicate that the complexity of the dinner meal data is higher than other meals. Thus, dinner may be considered the most at-risk component targeted as a priority to improve data quality.

The methodology proposed here offers a systematic approach to evaluating and improving dietary data quality in clinical trials. Greenberg et al [13] examined the outliers of daily energy and total fat intakes (determined as those three standard deviations from the mean) to assess dietary data quality. This method may be problematic as it could overlook errors existing within this range. The analysis applied to this study may provide a process model to conduct the assessment of dietary data entry errors. Furthermore, this study was more likely to provide the field with evidence related to the practice of dietary data entry. To further improve the operation management of dietary data generation.

There are limitations to conducting SDV on dietary intake data collected by an open-ended interviewer-administrated dietary assessment method, such as interviewer professional judgement related to training in nutrition and dietetics. Thus, performing SDV on the data set may also involve a degree of investigator subjectivity which can impact on the evaluation. Moreover, dietary data examined here was entered by a small group of qualified data entry personnel, hence investigating a larger group with differing levels of experience may identify further at-risk areas of dietary data quality.

3. Conclusion

The dinner meal appeared to be an at risk area of dietary data. The method developed in this analysis offers a systematic approach for dietary data improvement in the clinical research setting. Performing SDV on dinner meal data, particularly for quantity and frequency information may be a more efficient method to evaluate and improve dietary data quality at a larger scale.

References

[1] M.P. Ezzati and E.M.D. Riboli, Global Health: Behavioral and Dietary Risk Factors for Noncommunicable Diseases. *The New England Journal of Medicine* **369(10)** (2013), 954-64.

[2] F.E. Thompson, and A. Subar, Dietary assessment methodology, in *Nutrition in the Prevention and Treatment of Disease*, A.M. Coulston, C.J. Boushey, and M.G. Feruzzi, Editors. 2013, Elsevier;: Oxford, England. 5-46.

[3] D.G. Arts, N.F. De Keizer, and G.-J. Scheffer, Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *Journal of the American Medical Informatics Association* **9(6)** (2002), 600-611.

[4] M.L. Schuyl, and T. Engel, A Review of the Source Document Verification Process in Clinical Trials. *Drug Information Journal* **33(3)** (1999), 89-797.

[5] V. Tantsyura, et al., Risk-Based Source Data Verification Approaches: Pros and Cons. *Drug Information Journal* **44(6)** (2010), 745-756.

[6] L.C. Tapsell, et al., Interdisciplinary lifestyle intervention for weight management in a community population (HealthTrack study): Study design and baseline sample characteristics. *Contemporary Clinical Trials* **45, Part B** (2015), 394-403.

[7] Food Standards Australia New Zealand. NUTTAB Australian Food Composition Tables. *Food Standards Australia New Zealand*, (2010) Available from: .http://www.foodstandards.gov.au/science/monitoringnutrients/nutrientables/nuttab/Pages/default.aspx. Accessed May 13, 2015.

[8] M. Mealer et al., Remote Source Document Verification in Two National Clinical Trials Networks: A Pilot Study, *PLoS One,* **8** (2013), e81890

[9] J. Andersen et al., Impact of source data verification on data quality in clinical trials: an empirical post hoc analysis of three phase 3 randomized clinical trials. *British Journal of Clinical Pharmacology*, 2015. **79**(4): p. 660-668.

[10] L. Houston, Y. Probst, and A. Humphries. Measuring Data Quality Through a Source Data Verification Audit in a Clinical Research Setting. in Driving Reform: Digital Health is Everyone's Business: Selected Papers from the 23rd Australian National Health Informatics Conference (HIC 2015). *IOS Press* (2015), 107-113.

[11] L. Ward, et al., Data Entry Errors and Design for Model-Based Tight Glycemic Control in Critical Care. *Journal of Diabetes Science and Technology* **6** (2012), 135-143.

[12] D.R. Clarke, et al., Verification of data in congenital cardiac surgery. *Cardiology in the Young* **18(S2)**, (2008), 177-187.

[13] C. Greenberg, et al., Quality control of nutrient data entry for a long-term, multi-centre dietary intervention trial. *Journal of Food Composition and Analysis* **22, Supplement** (2009), S88-S92.