

Optimising the selection of food items for FFQs using Mixed Integer Linear Programming

Johanna C Gerdessen^{1,*}, Olga W Souverein², Pieter van 't Veer² and Jeanne HM de Vries²

¹Group Operations Research and Logistics, Wageningen University, Hollandseweg 1, 6706 KN Wageningen, the Netherlands; ²Division of Human Nutrition, Wageningen University, Wageningen, the Netherlands

Submitted 27 February 2013: Final revision received 13 November 2013: Accepted 13 November 2013: First published online 22 January 2014

Abstract

Objective: To support the selection of food items for FFQs in such a way that the amount of information on all relevant nutrients is maximised while the food list is as short as possible.

Design: Selection of the most informative food items to be included in FFQs was modelled as a Mixed Integer Linear Programming (MILP) model. The methodology was demonstrated for an FFQ with interest in energy, total protein, total fat, saturated fat, monounsaturated fat, polyunsaturated fat, total carbohydrates, mono- and disaccharides, dietary fibre and potassium.

Results: The food lists generated by the MILP model have good performance in terms of length, coverage and R^2 (explained variance) of all nutrients. MILP-generated food lists were 32–40% shorter than a benchmark food list, whereas their quality in terms of R^2 was similar to that of the benchmark.

Conclusions: The results suggest that the MILP model makes the selection process faster, more standardised and transparent, and is especially helpful in coping with multiple nutrients. The complexity of the method does not increase with increasing number of nutrients. The generated food lists appear either shorter or provide more information than a food list generated without the MILP model.

Keywords
Nutritional assessment
Methodology
Optimisation
Operations research
Multi-criteria decision making

FFQs are often used to assess the usual long-term dietary intake of individuals in nutritional epidemiological studies because they are easy to administer with relatively low costs⁽¹⁾. However, it is widely acknowledged that they cannot estimate the true usual intake of individuals without errors and that these errors affect the estimated diet–disease relationship. Nevertheless, FFQs may provide a more realistic instrument to assess long-term intake because they also capture infrequently consumed foods, whereas short-term instruments like 24 h recalls have presumably less bias but require many repeats. Indeed, a combination of FFQ with 24 h recalls was shown to provide a superior assessment compared with either method alone for some foods and nutrients⁽²⁾. However, practical and financial constraints still often favour the use of FFQs.

The basis of any FFQ is a food list enumerating all food items on which respondents are questioned. The aim in developing an FFQ is to select items for the food list such that as much information as possible is obtained for all nutrients of interest. However, the food list should not be too long in order to minimise the burden for respondents and the research costs. In the selection process decisions have to be taken on the level of aggregation of food items. Highly aggregated food items (such as 'fresh fruit') can capture a high coverage (i.e. the fraction of population

intake that is covered by the items in the food list) in relatively few items, but they often are not suitable to capture the between-person variance in intake⁽³⁾. Non-aggregated food items (single foods such as 'apples' or 'oranges') perform better with respect to capturing the between-person variance in intake. However, it takes many of these items to obtain sufficient coverage on all nutrients of interest.

To develop an FFQ, experts use standardised procedures. However, the process is time-consuming and the selection of food items depends strongly on the personal expertise of the expert. Molag *et al.*⁽⁴⁾ describe an automated procedure for selecting food items. In their selection procedure, one nutrient at a time is taken into account. For that nutrient, all food items are ranked based on their contribution to the variance in the population. Then the highest ranked food items are added to the food list. Selection stops as soon as the selected items suffice to obtain a predefined level of explained variance (R^2); for example, 80%. Then the same procedure is followed for the next nutrient, until all nutrients have been taken into account. The drawback of this procedure is that it only adds food items to the food list and never removes items. As most food items contribute to R^2 of several nutrients, it can be expected that food items added for one nutrient will also contribute to R^2 of another nutrient.

*Corresponding author: Email joke.vanlemmen@wur.nl

Thus the final food list will be unnecessarily long and the selection of food items will depend on the order in which nutrients are taken into account. Therefore this procedure does not suffice to minimise the number of questions in cases where the food list is targeted for multiple nutrients. Furthermore, the issue of choosing aggregation levels is not addressed. These drawbacks expose the need for a selection procedure that optimises the food list for multiple nutrients simultaneously, by taking into account the contribution of food items to both the level of intake and the variance of multiple nutrients. The procedure should also include the selection of food items at the best aggregation level. Selecting the subset of food items that maximises R^2 resembles the variable selection problem in statistics. However, the FFQ problem distinguishes itself from the common variable selection problem by the large number of available food items, the issue of selecting aggregation levels and the aim to optimise the food list for multiple nutrients simultaneously. The present paper describes how selection of food items for the food list of an FFQ can be modelled and supported by Mixed Integer Linear Programming (MILP) models. It also compares MILP-generated food lists with a food list developed with the procedure described in Molag *et al.*⁽⁴⁾ in terms of length, coverage and R^2 .

Methods

Data

To select food items to be included in the FFQ, food consumption data of the Dutch National Food Consumption

Survey of 1997/1998 of the 3524 individuals in the age group of 25–64 years⁽⁵⁾ were used. The food consumption was assessed using a 2 d food record, and converted into energy (index $n = 1$), total protein ($n = 2$), total fat ($n = 3$), saturated fat ($n = 4$), monounsaturated fat ($n = 5$), polyunsaturated fat ($n = 6$), total carbohydrates ($n = 7$), mono- and disaccharides ($n = 8$), dietary fibre ($n = 9$) and potassium ($n = 10$) with the Dutch food composition database of 1996⁽⁶⁾.

Aggregation level of food items: food tree

The food items are organised in a food tree with five levels, see Fig. 1. Level 5 contains all items that can be found as ‘food codes’ (single foods) in the NEVO food composition table⁽⁶⁾. Based on similarities in eating occasions, portion sizes and nutrient content, these detailed food items are aggregated into more aggregated food items and food groups⁽⁷⁾. The food groups at the highest aggregation level are the food groups as specified by NEVO⁽⁶⁾. The food tree contains 1697 items.

Measuring the quality of a food list: defining performance indicators

In the present paper we use three quantitative performance indicators to measure the quality of a food list: (i) *length* of the food list; (ii) *coverage* of the level of nutrient intake; and (iii) explained variance (R^2). The performance indicator *length* counts the number of food items in the food list (i.e. the number of selected ‘boxes’ in the tree of Fig. 1). *Coverage* describes the fraction of population intake covered by the items in the food list. The R^2 of a

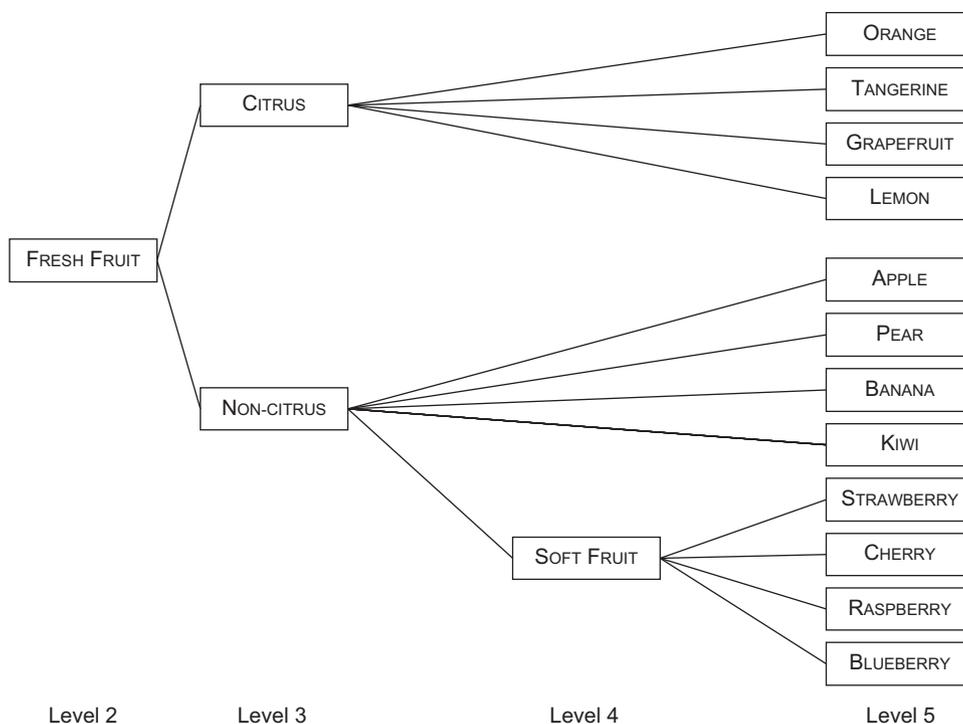


Fig. 1 Simplified and illustrative part of the tree structure that comprises FRESH FRUIT. Item FRESH FRUIT (level 2) can be further aggregated into item FRUIT (level 1), which also contains items such as canned fruits and dried fruits

nutrient is obtained from linear regression of total nutrient intake v . nutrient intakes of all food items in the food list; see online supplementary material A1.

FFQs that are targeted for multiple nutrients have a *coverage* and an R^2 for each nutrient n ($n = 1, \dots, N$), denoted as $coverage_n$ and R_n^2 , respectively.

Optimising the performance indicators of a food list

There is a crucial difference between *length* and $coverage_n$ of a food list, on the one hand, and R_n^2 , on the other hand: the contribution of each individual food item to *length* and $coverage_n$ of a food list can be uniquely quantified, whereas the contribution of an item to R_n^2 depends on the set of other items in the list. This makes it impossible to calculate in a straightforward way which combination of items provides maximal R_n^2 on all nutrients. Therefore, a three-step procedure was employed to select a set of items (i.e. a food list) with high R_n^2 .

1. A parameter $p_{j,n}$ is defined, which is a proxy for the contribution of item j to R_n^2 of a food list. P_n is defined as the sum of the $p_{j,n}$ of all selected items. It is therefore a proxy of the R_n^2 of the whole food list.
2. An MILP model is used to select the optimal set of items with respect to *length*, $coverage_n$ and P_n .
3. For the resulting set of items (i.e. food list), all R_n^2 are calculated.

The challenge is to find an effective proxy $p_{j,n}$, i.e. a proxy $p_{j,n}$ that demonstrates to be able to generate food lists with high R_n^2 . As it is expected that items with high intake and high variance in a population might be good candidates for a food list⁽³⁾, the following two implementations for proxy $p_{j,n}$ were tested.

1. Based on intake: $MOM1_{j,n}$ is the percentage that food item j contributes to the coverage of nutrient n intake of the population⁽⁸⁾. For items in level 5 we define $p_{j,n}$ as $MOM1_{j,n}$. For the aggregated items in the other levels we define $p_{j,n}$ as 90% of the sum of the $p_{j,n}$ in their constituent items. We refer to the intake-based proxy as $MOM1_{j,n}^*$. The 90% was chosen after experiments with several values (and calculating the R_n^2 of the resulting food lists), which showed that in general the value of 90% led to good results.
2. Based on variance: $MOM2_{j,n}$ is defined as the percentage that food item j contributes to sum of the variances of nutrient n within a specific level of the food tree⁽⁸⁾:

$$MOM2_{j,n} = \frac{\sum_{i=1}^I (F_{ij} - \bar{F}_{ij})^2}{\sum_{j=1}^J \sum_{i=1}^I (F_{ij} - \bar{F}_{ij})^2},$$

where $j = 1, \dots, J$ refers to all food items at a specific level in the food tree. We define $p_{j,n}$ as $MOM2_{j,n}$.

Mixed Integer Linear Programming model

The basis for the MILP model is the tree structure presented in Fig. 1. For every food item j in the tree we define a binary decision variable X_j , where:

$X_j = 1$ denotes that we decide to include item j in the food list

and

$X_j = 0$ denotes that we decide *not* to include item j in the food list.

We can express performance indicators *length*, $coverage_n$ and proxy P_n as linear functions of X_j :

$$length = \sum_{j=1}^J X_j,$$

$$coverage_n = \sum_{j=1}^J MOM1_{j,n} \cdot X_j,$$

for all nutrients $n = 1, \dots, N$

and

$$P_n = \sum_{j=1}^J P_{j,n} \cdot X_j,$$

for all nutrients $n = 1, \dots, N$.

Now we can formulate an MILP model that optimises one of these performance indicators while keeping the others at user-specified levels and ensuring that only feasible food lists are generated^(9,10). For example, if we want to optimise P_n while the *length* of the food list is no more than a pre-defined level (here: fifty items) and the $coverage_n$ of each nutrient is at least 75%, then an appropriate MILP model is:

Maximise

$$(1) \left\{ \sum_{n=1}^N P_n = \sum_{j=1}^J \sum_{n=1}^N P_{j,n} \cdot X_j \right\},$$

subject to the restrictions

$$(2) length = \sum_{j=1}^J X_j \leq 50,$$

$$(3) coverage_n = \sum_{j=1}^J MOM1_{j,n} X_j \geq 0.75,$$

for all nutrients $n = 1, \dots, N$

and

$$(4) \text{ the food list should not contain overlapping items.}$$

Restriction (4) is added in order to assure the model will generate only feasible food lists. For example, if item ORANGE is selected then CITRUS and FRESH FRUIT cannot be included, and if NON-CITRUS is included then FRESH FRUIT cannot be included, and neither can all items on the

right-hand side of NON-CITRUS. This can be modelled by adding one restriction for every item in level 5 of the food tree (see Fig. 1). For example, for ORANGE and CHERRY we add:

$$X_{\text{FRESH FRUIT}} + X_{\text{CITRUS}} + X_{\text{ORANGE}} \leq 1,$$

$$X_{\text{FRESH FRUIT}} + X_{\text{NON-CITRUS}} + X_{\text{SOFT FRUIT}} + X_{\text{CHERRY}} \leq 1.$$

Solving the model in equations (1)–(4) generates a feasible food list that has maximal value for average P_n among all food lists with at most fifty items and $\text{coverage}_n \geq 0.75$, provided that restrictions (2) and (3) are not conflicting. For example, if the user specifies that $\text{length} \leq 10$ and $\text{coverage}_n \geq 0.95$ then no food list is generated, because no such list exists.

An MILP-generated solution for the part of the food tree shown in Fig. 1 might be $X_{\text{CITRUS}} = X_{\text{APPLE}} = X_{\text{CHERRY}} = 1$, and $X_j = 0$ for all other food items. This should be interpreted as follows: three food items are included in the food list, namely CITRUS, APPLE and CHERRY. This implies that the model has chosen to aggregate ORANGE, GRAPEFRUIT, TANGERINE and LEMON into one aggregated item at level 3.

By interchanging the performance indicators in equations (1), (2) and (3), various models can be obtained. Several examples are provided in the online supplementary material A2.

Experiments

With the MILP model food lists have been generated of various lengths: 10, 20, ..., 150 items. Two different proxies $p_{j,n}$ were tested: $\text{MOM1}_{j,n}^*$ and $\text{MOM2}_{j,n}$. The quality of the resulting food lists was measured in terms of their performance indicators length , coverage_n and R_n^2 . Standard MILP-software (Xpress-Mosel 7.0.1) was used to solve the models (i.e. generate the food lists). Runtime of the MILP model was very small: for all instances a global optimal solution was found in less than 5 s.

Comparison with the ValNed

We compared length , coverage_n and R_n^2 of MILP-generated food lists with those of an actual FFQ, the so-called ValNed questionnaire⁽⁷⁾. This questionnaire was developed for the same nutrient set and with use of the same data source as the MILP-generated food lists. For constructing the food list of ValNed, the procedure of Molag *et al.*⁽⁴⁾ was used. The ValNed food list consisted of 117 items.

Results

Trade-off between length and R_n^2

For food lists of $\text{length} = 10, 20, \dots, 150$ items, all R_n^2 ($n = 1, \dots, N$) were calculated. Figure 2 shows the trade-off between length and R_n^2 . Figure 2 helps to weigh the

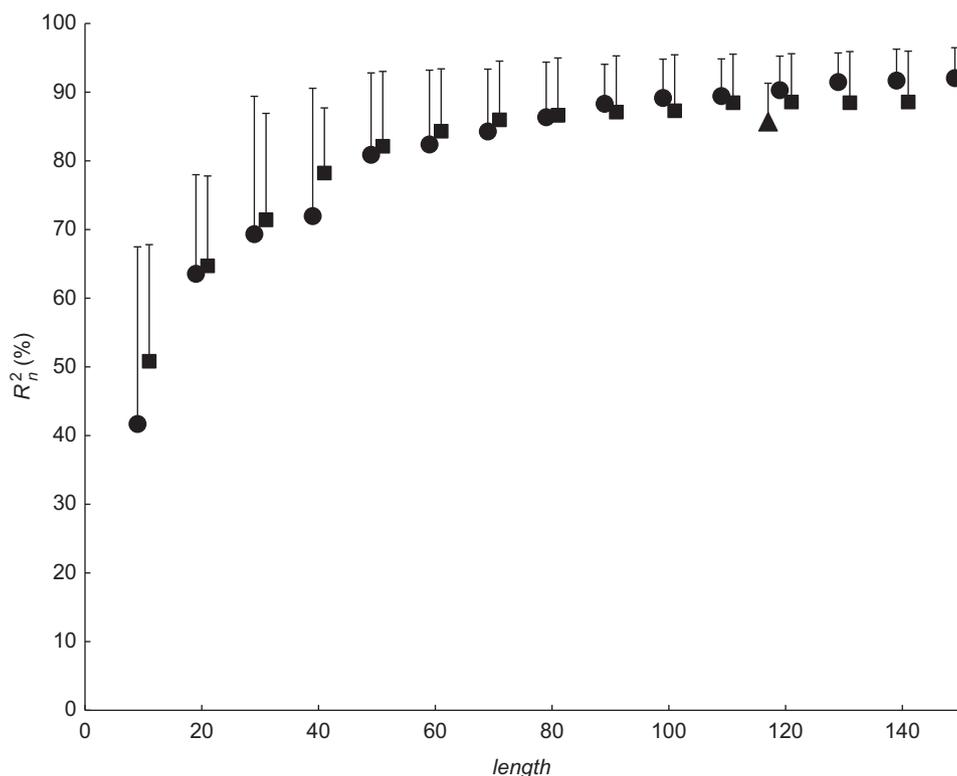


Fig. 2 Trade-off between the number of items in the food list (length) and R_n^2 (explained variance) for a coverage-based implementation and a variance-based implementation of proxy $p_{j,n}$ (food item $j = 1, \dots, J$; nutrient $n = 1, \dots, N$). For each food list the R_n^2 for all nutrients were calculated. The range of these R_n^2 is represented with a vertical bar. The lowest among the R_n^2 of a food list is represented by ● for $p_{j,n} = \text{MOM1}_{j,n}^*$ and by ■ for $p_{j,n} = \text{MOM2}_{j,n}$. As a point of reference, also the length and R_n^2 of ValNed (▲) are shown

amount of added information against the number of questions needed. For example, it took forty items to obtain a food list in which all nutrients have $R_n^2 \geq 70\%$, fifty items to have all $R_n^2 \geq 80\%$, and eighty items to have all $R_n^2 \geq 85\%$.

The choice for proxy $p_{j,n}$ (i.e. $MOM1_{j,n}^*$ or $MOM2_{j,n}$) had impact on the R_n^2 of the resulting food list. For food lists of up to seventy items, $p_{j,n} = MOM2_{j,n}$ was the best proxy. For food lists of more than ninety items, $p_{j,n} = MOM1_{j,n}^*$ was the best proxy.

Figure 3 shows the impact of *length* on the number of single food items selected for the FFQ. (All other selected items are aggregated items.) Both absolute and relative number of single food items grows with growing *length*, because a longer food list allows selection of more detailed food items and thus selection of relatively many single foods. ValNed uses fewer single foods than the MILP-generated food lists.

Comparison with ValNed

Length, $coverage_n$ and R_n^2 of the MILP food lists were compared with those of the food list of ValNed. Figure 2 shows that the MILP model obtained the same R_n^2 as ValNed in substantially fewer items, or vice versa obtained higher values for R_n^2 with the same number of items. This is further illustrated in Table 1, which shows the performance indicators of ValNed and of three food lists generated with the MILP model (with proxy $p_{j,n} = MOM1_{j,n}^*$ and the iterative improvement procedure as described in online supplementary material A2).

Food list MILP117 was generated with an MILP model that maximises P_n while putting an upper bound of 117 on the *length* of the generated food list. MILP117 has substantially higher values for $coverage_n$ and R_n^2 than ValNed, while *length* is the same. For generating MILP80 and MILP70, an upper bound of 80 and 70 respectively was put on the *length* of the food list. MILP80 has slightly

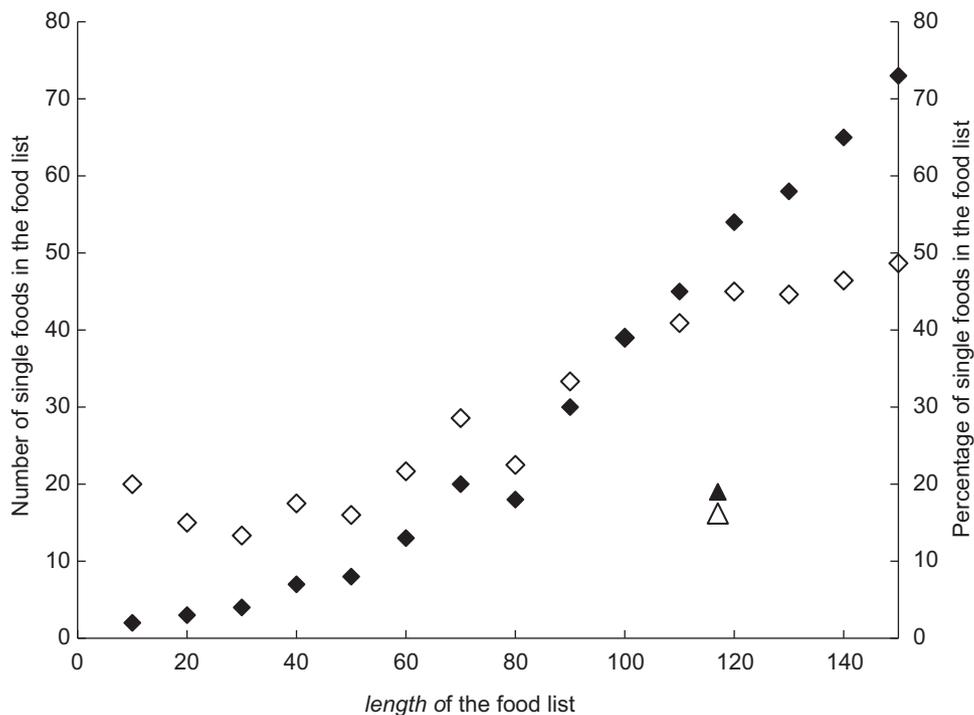


Fig. 3 Impact of the *length* of the food list on the absolute and relative number of single foods selected for the food list; ◆ represents the number of single foods in MILP food lists, ◇ the percentage of single foods in MILP food lists, ▲ the number of single foods in the ValNed food list and △ represents the percentage of single foods in the ValNed food list. For example, the MILP-generated food list of *length* twenty items contained three single foods (15%) and seventeen aggregated food items. The ValNed food list of *length* 117 items contained nineteen single foods (16%) and ninety-eight aggregated food items (MILP, Mixed Integer Linear Programming)

Table 1 Performance indicators of the food list of ValNed and three MILP-generated lists

Food list	<i>Length</i>	$Coverage_n$ (range, in %)	R_n^2 (range, in %)
ValNed	117	79.2–89.1	85.7–91.3
MILP117	117	96.0–97.3	90.5–95.2
MILP80	80	93.9–97.3	86.4–94.4
MILP70	70	91.4–97.2	84.3–93.4

MILP, Mixed Integer Linear Programming.

higher values for R_n^2 than ValNed, and MILP70 has slightly lower values for R_n^2 than ValNed. Both lists have substantially higher *coverage_n*. In other words, the MILP model generated lists that obtained the same R_n^2 as ValNed in substantially fewer (i.e. 32–40% less) items.

Discussion

The current paper presents a methodology for optimising food lists when developing FFQs. The decision problem of selecting food items for food lists was formulated as an MILP model with three performance indicators: *length*, *coverage* and R^2 . The MILP model generated food lists with good performance in terms of *length*, *coverage* and R^2 for all nutrients of interest. It supported the selection of the most informative aggregation level for food items and optimised multiple nutrients at the same time. The generated food lists were either shorter or provided more information than a food list generated without the MILP model.

The MILP model chooses the most informative combination of food items from different aggregation levels fast and objectively. Also, food lists of various *lengths* may be generated and the increase of *coverage* and R^2 obtained by adding more or other items to the food list may be investigated. The model provides objective information that can help to judge whether the extra information obtained by adding more food items justifies the additional burden for respondents and the additional research cost. With the MILP model multiple nutrients can be optimised simultaneously. In contrast with a manual selection procedure the number of nutrients has no impact on the complexity of the model. The results of the MILP-based selection procedure are highly reproducible. In addition, the MILP model can be included in a computer system.

In a typical MILP-supported selection procedure the MILP model is used to generate an initial food list, which is scrutinised by the nutritionist. The nutritionist indicates which constraints must be added to the model; for example, by specifying that some items should or should not be included in the food list to improve face validity. Then the MILP model is re-run in order to generate a food list that takes into account this expert knowledge. This loop is repeated until the nutritionist is satisfied. The nutritionist then decides how the items are ordered in the actual FFQ. In this iterative procedure the nutritionist is supported by the MILP model. It combines the strong points of human insight and experience on the one hand and the efficiency and accuracy of quantitative optimisation techniques on the other hand⁽¹¹⁾. It is complementary to the current practice of *post hoc* validation studies and it can help to guide efficient design of future web-based and personal monitoring tools.

Some considerations have to be taken into account.

It is important to realise that the food tree used here is constructed based on expert knowledge and that different

choices in the structure of the food tree would have led to different food lists.

Figure 2 shows the trade-off between *length* of the food list and R_n^2 for lists of ten to 150 items. The shortest of these lists were included for illustrative purposes; in practice they would not be used for covering the set of nutrients.

The data used to calculate the MOM1 and MOM2 values that are used as input are based on intake data of a Dutch adult population. Other populations will require other intake data, resulting in different food lists. The major advantage of the MILP model is that it generates new food lists fast and objectively when the user changes the input data, which facilitates easy adaptation to the characteristics and dietary habits of a population.

A limitation of the current data set is that only two subsequent food record days were available for each person. As result the between-person variance was artificially high since it contains part of the day-to-day variation within persons⁽¹²⁾. Also, the data set included multiple persons from the same household, lowering between-person variance in food intake and increasing correlations between foods. In general, the data used to generate the food list will have to be able to estimate the intake and variance of the foods and food groups adequately for the target population. Therefore, especially for infrequently consumed foods and when only one or a few days of recall per person are available, the sample size of the used survey will have to be large enough.

Table 1 indicates that for the used set of nutrients there is relatively little gain in either *coverage* or R^2 for food lists longer than seventy items. This might have been different for a different set of nutrients. If more (disperse) nutrients, e.g. vitamin C and carotenoids, were to be added the food list would probably need to be longer. The optimal length of the food list in general depends on the (number of) nutrients and on the dispersion of the nutrients through the available foods.

Even though these aspects may affect the resulting food lists, they have no effect on the methodology described here in terms of speed, transparency and reproducibility. This makes the MILP methodology useful for a large number of nutritionists worldwide who develop FFQs for a large variety of studies and different target populations^(13–15). It may help them to devise shorter questionnaires of which the performance is as good as that of FFQs with more food items, which generally provide the better results⁽¹⁶⁾. Validation with for instance biomarkers will have to answer the question on how well FFQs generated with this methodology perform in different populations.

Conclusion

MILP models can support development of food lists for FFQs. The results suggest that the MILP model makes the selection process faster, more standardised and transparent,

and is especially helpful in coping with multiple nutrients. The generated food lists appear either shorter or provide more information than a food list generated without the MILP model.

Acknowledgements

Sources of funding: This research received no grant from any funding agency in the public, commercial or not-for-profit sectors. *Conflicts of interest:* The authors declare that there are no conflicts of interest. *Ethics:* Ethical approval was not required. *Authors' contributions:* J.C.G. and P.v.t.V. conceived the concept of the study. J.C.G. designed and implemented the MILP model, conducted the MILP experiments and drafted the manuscript. O.W.S. performed the statistical analysis and helped to draft the manuscript. P.v.t.V. and J.H.M.d.V. supervised the research and gave critical comments on the manuscript.

Supplementary material

To view supplementary material for this article, please visit <http://dx.doi.org/10.1017/S1368980013003479>

References

1. Thompson FE & Byers T (1994) Dietary assessment resource manual. *J Nutr* **124**, 11 Suppl., S2245–S2317.
2. Carroll RJ, Midthune D, Subar AF *et al.* (2012) Taking advantage of the strengths of 2 different dietary assessment instruments to improve intake estimates for nutritional epidemiology. *Am J Epidemiol* **175**, 340–347.
3. Willett W (1998) *Nutritional Epidemiology*. New York: Oxford University Press.
4. Molag ML, de Vries JHM, Duif N *et al.* (2010) Selecting informative food items for compiling food-frequency questionnaires: comparison of procedures. *Br J Nutr* **104**, 446–456.
5. The Dutch Nutrition Centre (1998) *Zo eet Nederland: Resultaten van de Voedselconsumptiepeiling 1997–1998 (Results of the Dutch Food Consumption Survey 1997/1998)*. Den Haag: Voedingscentrum.
6. NEVO (1996) *Nederlands Voedingsmiddelentabel (Dutch Food Composition Table)*. Den Haag: De Commissie Nederlandse Voedingsmiddelentabel van de Voedingsraad.
7. Molag ML (2010) Towards transparent development of food frequency questionnaires. PhD Thesis, Wageningen University.
8. Mark SD, Thomas DG & Decarli A (1996) Measurement of exposure to nutrients: an approach to the selection of informative foods. *Am J Epidemiol* **143**, 514–521.
9. Gerdessen JC, Slegers PM, Souverein OW *et al.* (2012) Use of OR to design food frequency questionnaires in nutritional epidemiology. *Oper Res Health Care* **1**, 30–33.
10. Gerdessen JC, Claassen GDH & Banasik A (2013) General 0–1 fractional programming with conditional fractional terms for design of food frequency questionnaires. *Oper Res Lett* **41**, 7–11.
11. Claassen GDH, Hendriks THB & Hendrix EMT (2007) *Decision Science: Theory and Applications*. Wageningen: Wageningen Academic Publishers.
12. Lambe J, Kearney J, Leclercq C *et al.* (2000) The influence of survey duration on estimates of food intakes and its relevance for public health nutrition and food safety issues. *Eur J Clin Nutr* **54**, 166–173.
13. Cade J, Thompson R, Burley V *et al.* (2002) Development, validation and utilisation of food-frequency questionnaires – a review. *Public Health Nutr* **5**, 567–587.
14. Bharathi AV, Kurpad AV, Thomas T *et al.* (2008) Development of food frequency questionnaires and a nutrient database for the prospective urban and rural epidemiological (PURE) pilot study in South India: methodological issues. *Asia Pac J Clin Nutr* **17**, 178–185.
15. Wakai K (2009) A review of food frequency questionnaires developed and validated in Japan. *J Epidemiol* **19**, 1–11.
16. Molag ML, de Vries JHM, Ocké MC *et al.* (2007) Design characteristics of food frequency questionnaires in relation to their validity. *Am J Epidemiol* **166**, 1468–1478.