Food Control 47 (2015) 212–220

Contents lists available at ScienceDirect

Food Control

journal homepage: www.elsevier.com/locate/foodcont

A Bayesian network to optimise sample size for food allergen monitoring

C.F. Elegbede*, Alexandra Papadopoulos, Julie Gauvreau, Amélie Crépet

Risk Assessment Department (DER), French Agency for Food, Environmental and Occupational Health & Safety (ANSES), Maisons-Alfort, France

ARTICLE INFO

Article history:
Received 14 December 2013
Received in revised form
19 June 2014
Accepted 20 June 2014
Available online 28 June 2014

Keywords:
Sample size optimisation
Bayesian modelling
Peanut allergens
Labelling

ABSTRACT

Generally, sampling size is optimised considering a single specific constraint. However, for financial reasons, only one sample is usually defined and used to satisfy several objectives. It is therefore crucial to choose a sample that meets all the required objectives.

This paper proposes an original method for optimising a sample plan to monitor allergen traces in products consumed by allergy sufferers. The proposed method, based on a Bayesian network, enables several different constraints to be considered within a single model and the integration of literature data on concentration levels of allergen traces in food. Moreover, the construction of a three-stage sampling plan took into account the consumption preferences of peanut allergy sufferers between products with or without labels on the presence of allergen traces, and between the categories and subcategories of products. This method was applied to data from the MIRABEL project which aims to assess risks related to peanut traces for French allergy sufferers.

The results show how the model used all the available information and constraints to balance the total number of samples set at 900 for food categories/subcategories and labelling types. As required, the model favoured the most consumed product categories and subcategories. At the same time, it increased the number of samples when peanut concentration is low. This helps reduce the uncertainty on peanut concentrations in these products and consequently on risk estimation.

In conclusion, the proposed method is a useful tool for public administrations, risk assessors and risk managers to improve sampling plans for monitoring allergen traces or other health hazards in food.

1. Introduction

Determining optimal sample size is one of the most recurrent issues when designing a sampling plan. A small sample cannot assure satisfactory statistical results, while too large a sample may be unnecessary and complicated to use because of financial and time constraints. Sample size is therefore driven by several constraints such as research objectives, type of data, or confidence in the estimates. In practice, sample size is generally optimised on the basis of a single constraint. For instance, Adcock (1993) discussed a procedure for calculating sample sizes for multinomial sampling. Weiss (1997) presented approaches for optimising sample size based on different types of confidence interval for a binomial success probability. Furthermore, any powerful comparative statistical tests may be another useful consideration for optimising sample size. For instance, DasGupta and Vidakovic (1997) discussed the sample size problem for balanced one-way ANOVA. Moreover, a single sample is used to meet the specific objectives of those studies. It is therefore not easy to choose an appropriate sample that fulfils all the specific objectives.

This paper proposes an original approach for optimising sample size by combining different data under several constraints. The approach was specifically developed for food sampling to monitor allergen traces in food consumed by the allergic population. Implicitly, analysing allergen traces aims to: (i) assess the probability of finding allergen traces, (ii) assess concentration levels, and (iii) compare the probability and concentration of allergen traces in the different products or categories/subcategories of products. Indeed, regarding peanut traces, food products which may contain traces can be divided into different categories (e.g. cereals, chocolates, biscuits, etc.) and subcategories depending on the ingredients or the flavour (cereals containing chocolate or fruit etc.). These products contain different amounts of peanut traces (Rimbaud, Heraud, La

* Corresponding author. ANSES 27-31 avenue du Général Leclerc, MAISONS-ALFORT, 94701, France. Tel.: +33 1 56 29 52 04; fax: +33 1 49 77 38 92.
E-mail address: chabi.elegbede@gmail.com (C.F. Elegbede).

http://dx.doi.org/10.1016/j.foodcont.2014.06.039
0956-7135/

Vieille, Leblanc, & Crepet, 2013). Moreover, they may be unequally consumed by allergy sufferers. Therefore, to ensure their satisfactory representation in the overall sample and to make comparison possible, the number of food products has to be determined when optimising the sampling size per category and subcategory.

This research is part of the MIRABEL science project (Model Integrated Risk for Allergy, Bayesian Estimation for Life quality; Guenard-Bilbault et al., 2012) designed to assess the risk to French allergy sufferers from the unintended presence of peanut traces in food. Food allergies are a public health concern, affecting nearly 2 million people in France. The most prevalent is an allergy to peanuts, accounting for 0.3%–0.75% of the French population (Moneret-Vautrin, 2008; Rancé, Grandmottet, & Grandjean, 2005). The unintended presence of peanut traces in food is involved in 8.6% of serious allergic reactions (ANSES, 2008). The risks related to this allergen depend on different factors: food consumption (probability and quantity), allergen presence in food (probability and concentration) and reaction threshold (Rimbaud, Heraud, La Vieille, Leblanc, & Crepet, 2010; Spanjersberg, Kruizinga, Rennen, & Houben, 2007). The MIRABEL project conducts field surveys to fill in the current data gaps for each component and propose statistical developments to improve allergen risk assessment. Since no data on the probability and concentration levels of unintended peanut traces in food is available for the French market, one of MIRABEL's main objectives is to collect relevant data. Up to now, none of the European or American studies conducted to investigate the level of unintended allergen traces in food products has defined an optimal food sampling plan. Indeed, current studies consider neither previous knowledge on unintended peanut traces in food, neither the consumption behaviour of the allergic population nor their habits with respect to precautionary labelling. However, precautionary labels such as “may contain traces of peanuts” are often used to warn allergy sufferers about the potential unintended presence of allergens in food products. Hefle et al. (2007) showed that labelled (with precautionary labels related to allergen traces on package) and unlabelled products (without precautionary labels related to allergen traces on package) may be consumed in unequal quantities by allergy sufferers. Moreover, labelled and unlabelled products may have different peanut trace levels. Therefore, an optimal sample size should considerer allergic consumer behaviours toward labelling and must be defined per labelling type (labelled/unlabelled).

In this work, a Bayesian network is proposed to optimise the sample size by combining different data under several constraints. Sample size optimisation approaches have been developed in both frequentist (Lachin, 1981) and Bayesian contexts (Adcock, 1988, 1997; Pham-Gia & Turkkan, 1992; Sahu & Smith, 2006). Nevertheless, Bayesian methods are becoming more popular because of their capacity to integrate prior information and propose efficient algorithms to estimate parameter distributions for complex models. Indeed, when optimising sample size, it is sometimes necessary to assign numerical values to model parameters. These values can be estimated from the literature, previous data or expert knowledge. The main limitation of a frequentist method is that only a point estimate is assigned to the parameters. These methods do not therefore take into account the epistemic uncertainty (i.e. lack of knowledge) of model parameters. In contrast, Bayesian methods take into account available information by using prior distributions for model parameters, rather than a point estimate, so uncertainty is taken into account in the results.

The next section contains the specific constraints defined to optimise food sample size in the context of allergy risk assessment. It also describes the data, sampling plan and Bayesian network developed. The last section presents and discusses the proposed method and results.

2. Materials and methods

2.1. Constraints for optimising food sample size

The sample size was optimised so that the final sample: (i) was composed of two labelling types (labelled/unlabelled) and several food product categories and subcategories regarding the designed sampling plan, (ii) reflected data from the literature on peanut trace probabilities and concentrations (iii) reflected MIRABEL data on the food product consumption probabilities by peanut allergy sufferers and (iv) made all comparative statistical tests powerful.

2.2. Data

2.2.1. MIRABEL programme

The MIRABEL project was divided into four complementary and interactive tasks. The first task was to conduct a survey among peanut allergy sufferers to collect data on their food consumption, reaction thresholds and their willingness to pay for new food products with labels on allergen concentration limits. Peanut allergy sufferers were recruited by their allergist, members of the French Allergyvigilance network, during a consultation between March 2012 and December 2013. The challenge of the second task was to analyse the unintended presence of allergen in products listed in the consumer behaviour survey. The aim was to quantify the probability and concentration of peanut traces in the 900 most-consumed products by MIRABEL subjects. Task 3 was devoted to developing a probabilistic risk model, applying it to the data collected during the previous tasks and testing different risk management options. This task, integrating all the knowledge acquired in the other tasks, was the project's main thread. The objective of task 4 was to provide a cost-benefit analysis of possible regulatory options such as new labelling based on the results of the consumer survey for peanut allergy sufferers. This analysis was designed to help assess the impact of *ex ante* regulatory measures before their actual implementation.

2.2.2. Choice of categories and subcategories of food products

This task was carried out at the beginning of the MIRABEL project to identify the generic food to focus on. The choice depended on two criteria: (i) products for which the literature revealed an unintended presence of peanut traces; (ii) products bearing a precautionary label on their packaging. This information was available from the French Observatory of Food Quality database which centralises, at the branded product level, all data provided on labels (such as the ingredient list and precautionary labels) to monitor processed food quality, over time (OQALI, Menard, et al. (2011)). The considered information was collected between 2008 and 2010.

Ten categories of food product were selected and divided into subcategories depending on the ingredients or flavour. Table 1 lists the food product categories and subcategories.

2.2.3. Behaviour of allergic consumers

For the first task of the MIRABEL project, peanut allergy sufferers were asked about their consumption of products in the selected food categories and subcategories. The allergic individuals self-reported their consumption frequency per day, week, month or year. The quantities consumed were also recorded by means of photographs. The products were precisely described, and a record made of their name, brand, flavour and labelling type (labelled/unlabelled). Table 1 presents the consumption probabilities of each category $Cons_{Cat,i}$ and subcategory $Cons_{SCat,i,j}$ per labelling type calculated from the responses of 194 individuals. The population was mainly composed of children (92%) aged between 3 and 16