MCRA 7

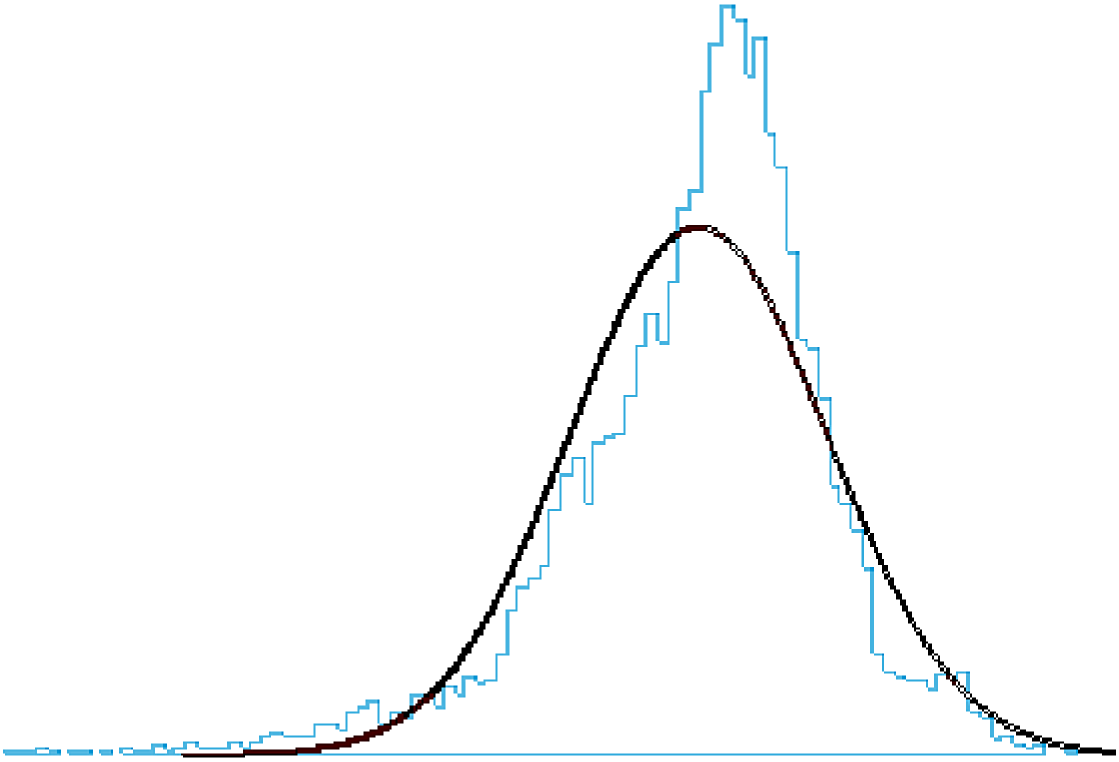
a web-based program for Monte Carlo Risk Assessment

Reference Manual 2011-12-19

documenting MCRA Release 7.1

Waldo J. de Boer
Hilko van der Voet

Contributions to MCRA were made by: Paul W. Goedhart, Gerie W.A.M. van der Heijden, Sanne Korzec, Helen Owen (Food and Environment Research Agency, UK), Jac T.N.M. Thissen



Biometris
Wageningen University and Research centre
RIVM
Centre for Substances and Integrated Risk Assessment,
National Institute for Public Health and the Environment

19 December 2011

Biometris is the unit for Mathematical and Statistical Methods of Wageningen University & Research centre

Post address
P.O. Box 100
6700 AC Wageningen
The Netherlands
Visiting address
Droevendaalsesteeg 1, building no. 107
6708 PB Wageningen
Telephone: +31 (0)317 484085
Telefax: +31 (0)317 483554

RIVM National Institute for Public Health and the Environment

Post address
P.O. Box 1
3729 BA Bilthoven
The Netherlands
Visiting address
Antonie van Leeuwenhoeklaan 9
3721 MA Bilthoven
Telephone: 0031 30 2749111
Telefax: 0031 30 2742971

1 Introduction to MCRA

MCRA (Monte Carlo Risk Assessment) is a computational tool for dietary risk assessment of substances in foods based on monitoring data concerning the quality of foods and agricultural products. Intake (exposure) assessment is an important step in risk assessment of substances found in food, such as agricultural chemicals (*e.g.* pesticides, veterinary drugs), toxins (*e.g.* mycotoxins), environmental contaminants (*e.g.* dioxins), nutrients and micronutrients (*e.g.* fat, vitamins).

1.1 Model description

This manual describes the stochastic (or Monte Carlo) models behind the MCRA program. These models assess acute (short-term) or chronic (long-term) risks due to dietary intake by combining food consumption survey data and concentration data from *e.g.* monitoring programs.

Food consumption data may arise from different sources. Typically, national food consumption surveys or monitoring programs provide information on food intake in the general population. For example, from the Dutch Food Consumption Survey (1997) food consumption patterns ($x_1, \ldots, x_p$), body weight (w), age (a) and sex (s) are available for 6250 individuals on 2 consecutive days. When concentrations are not measured on consumed foods, a composition database is necessary to convert the amounts of food as consumed (*e.g.* pizza) to amounts of foods as measured ($x_1, \ldots, x_p$) which are used in the model. Van Dooren *et al.* (1995) provide such a conversion for the Dutch situation. Concentration data may be available from different sources. In some countries national monitoring databases exist, which are useful for the risk assessment of substances already in use. For example, the Dutch KAP database (van Klaveren 1999) stores annually more than 200,000 records of measurements originating from food monitoring programs for meat, fish, dairy products, vegetables and fruit.

Basically, MCRA simulates daily consumptions by sampling a food consumption database and combines these with a random sample from either a concentration database (empirical distribution) or a parametric distribution of concentrations. The result is a full *distribution* of intakes, rather than traditional deterministic methods which only provide a point estimate. Percentiles of the intake distribution can be used to assess risks by relating them to *e.g.* an acute reference dose (ARfD).

The basic model for the intake of a special substance in an acute risk analysis is:

$$y_{ij} = \frac{\sum_{k=1}^{p} x_{ijk} \; c_{ijk}}{w_i}$$

where y_{ij} is the intake by individual i on day j (in microgram substance per kg body weight), x_{ijk} is the consumption by individual i on day j of food k (in g), c_{ijk} is the concentration of that substance in food k eaten by individual i on day j (in mg/kg, 'ppm'), and w_i is the body weight of individual i (in kg). Finally, p is the number of foods accounted for in the model.
Note that the definition of 'food' is flexible: it may represent a Raw Agricultural Commodity (RAC), *e.g.* 'apple', but it may also specify subdivisions, *e.g.* 'apple, peeled' or 'apple, imported'.

The quantities x_{ijk}, w_i and c_{ijk} are assumed to arise from probability distributions describing the variability for food consumption and weight, $p(x_1,\ldots,x_p,w)$, and for concentrations of substances in each food, $p_k(c)$. In principle, these probability distributions may be parametric (*e.g.* completely defined by the specification of some parameter values) or empirical (*e.g.* only implicitly defined by the availability of a representative sample). Given these probability distributions (or estimates thereof) MC-simulations can be used to generate an estimate of the probability distribution $p(y_{ij})$ to assess acute risks by intake of the substance.

In a chronic exposure assessment, the main interest goes to the fraction of individuals with a usual intake per day higher than an intake limit. Usual intake is defined here as the long-run average of daily intakes of a substance by a individual. Usually, food consumption data are available for individuals on 2 (or more) consecutive days. We assume an equal number of days for each individual. This is in confirmity with our method of data entry for consumption. As a consequence, days without consumptions do have zero intake. MCRA calculates the distribution of the usual intakes over individuals based on the average concentration and the empirical distribution of intake between individuals and between different intake days of the same individuals. Percentiles of this usual intake distribution can then be related to *e.g.* the acceptable daily intake (ADI).

The basic model for the intake in a chronic risk analysis is:

$$y_{ij} = \frac{\sum_{k=1}^{p} x_{ijk} \ c_k}{w_i}$$

where y_{ij}, x_{ijk} and w_i are defined as before but now concentrations of the substance found in food k enter the model as an average of all values, c_k .

In MCRA 7.1 the following models are available to assess chronic risks:

1) the betabinomial-normal (BBN) model (see 5),
2) the logisticnormal-normal model, in a full version that includes the estimation of correlation between intake frequency and amount (LNN), and in a simpler version without this estimation (LNN0) (see 6)
3) the discrete/semi-parametric model known as the Iowa State University Foods (ISUF) model; (see 7).
4) the observed individual means (OIM) model (see 8)

The models for acute and chronic risks allow for effects of food processing between monitoring and ingestion, they use information on Limit of Reporting (LOR) and percent crop treated to check whether non-detects present a source of uncertainty. For acute risks, unit variability either from available data or using default assumptions can be modelled. Uncertainty of percentiles or intake limits can be established by resampling methods.

Depending on the problem, MC-samples may be drawn from the complete database, from a day- or age-restricted subset or from consumption-days only. In some cases there is insufficient information for specific subgroups in the population. For example, in a study on infants (age up to 12 months), a separately constructed food consumption database has been used (Boon *et al.* 2003).

1.2 Data needed

The data needed for MCRA are stored in MS Access databases or Excel. The database format requires understanding about building up a relational database using the primary sources of the data. In exchange, flexibility to pre-process the data is offered and results may be investigated in greater detail. The Excel type of data do have a simple two-way data lay-out. Find in **MCRA 7 Data Formats** a full description of how to prepare the data.

Basically, input data for MCRA usually originate from two sources: food consumption surveys and monitoring programs on substances found in foods. Often, additional tables are needed to link consumption data to concentration data or to implement model options like unit variability or processing. Figure 1 presents the linkage between tables: consumption data are linked directly to concentration data or in an indirect way, through the use of food composition data, food marketshare data, processing data or by the use of a supertype algorithm.

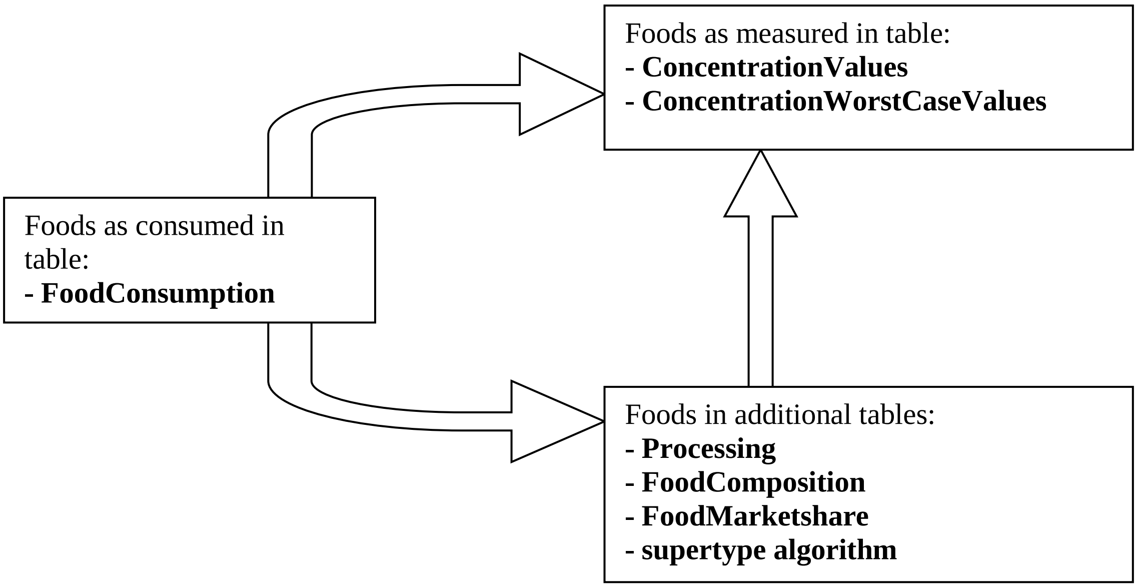

Foods as measured in table:
- ConcentrationValues
- ConcentrationWorstCaseValues
Foods as consumed in table:
- FoodConsumption
Foods in additional tables:
- Processing
- FoodComposition
- FoodMarketshare
- supertype algorithm

Figure 1: Links between consumption and concentration data

Consumption data are consumed portions of food (consumed at different days) of individuals. To get standardized intakes, supply the weight of each individual. Other characteristics of the individuals, like age and/or sex, may be used in further analyses.
Concentration data on substances are the amounts of the substance found on monitoring samples of food.
The category additional tables provide information that links consumption data to concentration data or store information for more sophisticated analyses like unit variability (see Figure 1). Food composition data specifies the composition of foods. So, speaking about pizza, the composition specifies proportions for *e.g.* wheat, tomato, cheese etc. Food marketshare data specifies the proportion of subtypes, so for apple, marketshares are defined for *e.g.* Jonagold, Granny Smith, Golden Delicious etc. Processing data specify the unprocessed food, the processed food and the corresponding processing factors, *e.g.* for grapes raisins are specified. The supertype of a food is, if needed, automatically determined. So the supertype of *e.g.* Granny Smith is apple.

The MCRA system has a central database with example data. However, MCRA is primarily designed to work with user databases, or with a mixture of user and centrally supplied data. For example, provide your own data on concentration levels and combine these with the centrally supplied consumption data. Be careful when combining tables from different databases: codes of foods of centrally supplied data and your own data should be consistent.

1.3 Get started

To use MCRA, go to https://mcra.rivm.nl. As a potential new user, first fill in the registration form. After login, the central menu is entered and from here all tasks and corresponding actions are started.

The central menu (Figure 2) contains four main tasks which are described as:

- Data Selection (Access [mdb], Excel [xls] or Simulated Data [xls])
- Specify Model (specification of input options)
- Set and Run (specification of output options, start Monte Carlo Risk Assessment)
- View Output (managing output)

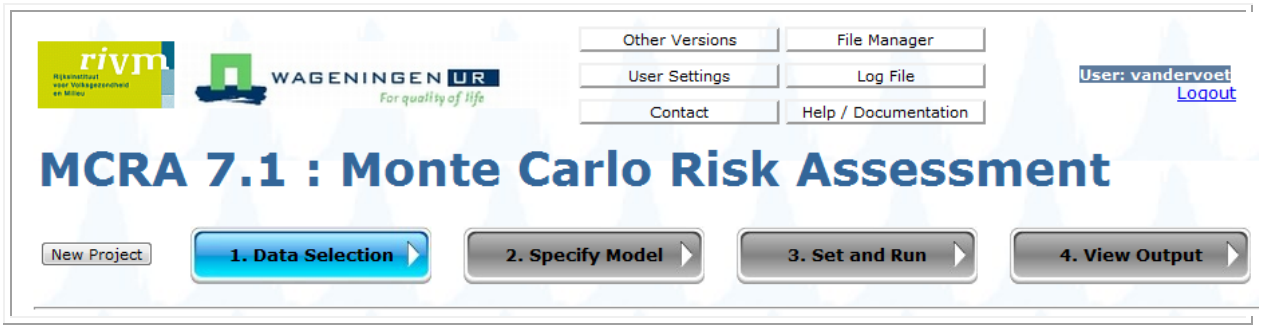


Figure 2: Central menu

A main task is started by clicking the button. Then, a menu containing actions related to the main task is displayed. After clicking a main button, it turns into blue to indicate that the task is active. For a first time user, the figure above shows the central menu and Data Selection can be started (only available option). Otherwise, press New Project to clear all selections.

For a short introduction in MCRA, we refer to **MCRA 7 Examples** or **MCRA 7 Overview**.

2 Acute risk assessment

Substance concentrations in the various foods are independent and therefore can be modelled by univariate distributions.

2.1 Empirical modeling of concentrations

In the empirical (non-parametric) approach, concentrations are sampled at random from the available data and combined with the consumption data to generate a new distribution of intake values. To assess the risk-intake, percentiles of the intake distribution are estimated.

2.2 Parametric modeling of detects and nondetetects

In the parametric approach, concentrations per food are sampled from parametric distributions. A special feature of concentration data is that the large majority of measured concentrations (often more than 80%) is recorded as zero (non-detects). These values may correspond to true zero concentrations (for example because the substance is never used in the specific food), or they may correspond to low concentrations which are below a pre-established Limit of Reporting (LOR).

Regarding all non-detects as censored data values is not always valid. Alternative models exists to fit data that contain non-detects. In [EFSA, draft interim scientific report, 2009], a review is given on the most commonly used statistical methods to deal with non-detects. Among them are substitution, log-probit regression, maximum likelihood estimation and non parametric methods. In the draft EFSA report, the question whether the non-detects are true zero's or low concentrations is not considered, and only described in terms of a combination of more than one log-normal distribution, e.g. binomial and a lognormal.

The lognormal distribution (logarithmic transformed values) with parameters μ and σ^2 has been selected as being both theoretically sensible and practically useful (Shimizu & Crow 1988, van der Voet *et al*. 1999). Based on this priciple, we then have the following six methods:

1. empirical (nLor ≥ 1)
2. mixture of non-detect spike and lognormal (nLor ≥ 1)
3. mixture of non-detect spike and truncated lognormal (nLor = 1)
4. censored lognormal (nLor ≥ 1)
5. censored lognormal with estimated LOR (nLor = 1)
6. mixture of zero spike and censored lognormal (nLor ≥ 1)

with nLor indicating whether multiple values for the LOR are allowed or not.

Additional options for the first three models are:

- if there are non-detect data, these can be replaced by *f* x LOR for a specified value *f*,
- if *f* > 0, an additional option is to apply percentage crop treated to force true zero concentrations for part of the non-detect data. For legal applications of substances like pesticides, data may be available about the percentage of the crop which receives treatment. When a substance can enter the food chain only via crop treatment, and when the percentage of crop treated is (approximately) known to be $100p_{crop\text{-}treated}$, then this knowledge may be used to infer that $100(1\text{-}p_{crop\text{-}treated})$% of the monitoring measurements should be real zeroes, contributing nothing to intake of the substance, whereas other non-detects in the monitoring data could have any value below the LOR.

For $100(p_{non\text{-}detect} + p_{crop\text{-}treated} - 100)$% of the monitoring measurements, 0 and LOR represent best-case and worst-case estimates. A simple way (tier 1 approach) to consider the uncertainty associated with non-detects is to compare intake distributions for these best-case and worst-case situations.

Method 2 (mixture of non-detect spike and lognormal) can be defended if all positive values are assumed to be above LOR, and P(c<LOR) would be very small for the fitted distribution. It would then be logical to apply *f* = 0 for the nondetects.
Method 3 (mixture of non-detect spike and truncated lognormal) estimates a truncated lognormal distribution. The estimated P(c<LOR) should be lower than the fraction nondetects, and for the difference *f* = 0 would be a logical choice.
Methods 4 (censored lognormal) and 5 (censored lognormal with estimated LOR) assume that there are no true zeroes, which might be a reasonable assumption for many contaminants, though not for artificially added substances. With model 5 the reasonableness of the given LOR value can be checked (assuming the lognormal model).
Method 6 (mixture of zero spike and censored lognormal) fits a mixture distribution, where the nondetects are divided over a spike of true zeroes and the censored tail of the lognormal distribution.
Method 1 (empirical) is the parameter free alternative (default) and samples concentration values directly from the empirical concentration distribution using both detect and nondetect data. It requires to specify a value *f* for the nondetects (also true for methods 2 and 3). This approach requires more data to obtain a satisfying representation of the full distribution.

2.3 Estimation

The parameters of the truncated lognormal, censored lognormal and censored lognormal mixture model may be estimated using maximum likelihood. The censored lognormal with estimated LOR is an iterated version of the censored lognormal and searches for the best value of the LOR under the assumption that the observed fraction of non-detects equals the predicted fraction of non-detects. This often improves the fit of the data and supports the notion that values of the LOR are not precisely reported by the analytical labs.

Let *x* denote a random variable from a lognormal distribution. Then, the log transformed variable *y* = ln(*x*) is normally distributed with mean μ_y and variance σ_y^2.
The probability density function (p.d.f.) of *y* may be expressed as:

$$f_y(y; p_0, \mu_y, \sigma_y^2) = p_0 I(y;0) + \{(1-p_0)(1-I(y;0)\} * \frac{1}{\sqrt{2\pi}\sigma_y} \exp(-\frac{(y-\mu_y)^2}{2\sigma_y^2})$$

where $p_0 = \Pr(y < \log(x_{lor})$, x_{lor} is the limit of reporting and I(*y*;0) is an indicator function for $y < \log(x_{lor})$. For $p_0 = 0$, the p.d.f. of *y* reduces to the usual lognormal density.

The left truncated density for $y \geq \log(x_{lor})$ equals (assume one LOR):

$$f_y(y; \mu_y, \sigma_y^2) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp(-\frac{(y-\mu_y)^2}{2\sigma_y^2}) / (1-\Phi(z))$$

with $\Phi(.)$ the standard normal c.d.f. and $z = (\log(x_{lor}) - \mu_y) / \sigma_y$

Model parameters are estimated using maximum likelihood estimation based on the loglikelihood functions specified below:

1) mixture of zero spike and censored lognormal:

$$\log L(p_0, \mu_y, \sigma_y^2) = \sum_{i=1}^{n_0} \log(p_0 + (1-p_0)\Phi(z_i)) + n_1 \log(\frac{1-p_0}{\sqrt{2\pi}\sigma_y}) - \sum_{i=n_0+1}^{n} \frac{(y_i-\mu_y)^2}{2\sigma_y^2}$$

4) censored lognormal

When $p_0 = 0$, loglikelihood (1) reduces to

$$\log L(\mu_y,\sigma_y^2) = \sum_{i=1}^{n_0}\log(\Phi(z_i)) + n_1\log(\frac{1}{\sqrt{2\pi}\sigma_y}) - \sum_{i=n_0+1}^{n}\frac{(y_i-\mu_y)^2}{2\sigma_y^2}$$

and for

5) censored lognormal with estimated LOR (assume one LOR)

$$\log L(lor,\mu_y,\sigma_y^2) = n_1\log(\Phi(z_{lor})) + n_1\log(\frac{1}{\sqrt{2\pi}\sigma_y}) - \sum_{i=n_0+1}^{n}\frac{(y_i-\mu_y)^2}{2\sigma_y^2}$$

3) mixture of non-detect spike and truncated lognormal (assume one LOR)

Ignoring the n_0 values below x_{lor} leads to:

$$\log L(\mu_y,\sigma_y^2) = -n_1\log(1-\Phi(z)) + n_1\log(\frac{1}{\sqrt{2\pi}\sigma_y}) - \sum_{i=n_0+1}^{n}\frac{(y_i-\mu_y)^2}{2\sigma_y^2}$$

2) mixture of non-detect spike and lognormal

Ignoring the truncated part leads to:

$$\log L(\mu_y,\sigma_y^2) = n_1\log(\frac{1}{\sqrt{2\pi}\sigma_y}) - \sum_{i=n_0+1}^{n}\frac{(y_i-\mu_y)^2}{2\sigma_y^2}$$

where $y_i = \log(x_i)$, $\Phi(.)$ is the standard normal c.d.f., $z = (\log(x_{i,lor}) - \mu_y)/\sigma_y$, $z_{lor} = (\log(lor) - \mu_y)/\sigma_y$, n_0 number of censored values ($x_i < x_{i,lor}$), n_1 number of uncensored values ($x_i >= x_{i,l\ or}$) and x_i, $i = 1\ldots n$
The loglikelihood functions are evaluated in R, using the *optim* algorithm to find estimates for μ_y, σ_y^2 and p_0.

In the basic model, for method 1, 2 and 3 we have:

$$c_{ijk} = I_{ijk}\cdot cpos_{ijk}$$

with I_{ijk} indicating whether a concentration is sampled (I_{ijk} =1) or not (I_{ijk} =0), and $cpos_{ijk}$ is the concentration value according to the chosen method. The probability of I_{ijk} being 1 or 0 depends on the number of detects found for food *k* and I_{ijk} is sampled separately for each individual *i* on occasion *j*.
For method 4 and 5, $p_0 = 0$ and the basic model reduces to:

$$c_{ijk} = cpos_{ijk}$$

For method 6, where $p_0 > 0$, $I_{ijk} = p_0$ and c_{ijk} is sampled as in method 1, 2 and 3.

Occasionally, estimation of the model parameters (mean, variance and zero spike) may fail because concentration data on specific foods are sparse or even missing.

3 Modelling processing effects

Concentrations in the consumed food may be different from concentrations in the food as measured in monitoring programs (typically raw food) due to processing, such as peeling, washing, cooking etc. In general, we assume the model:

$$cpos_{ijk} = f_k \cdot c_{ijk}$$

where c_{ijk} is the concentration in the raw food, and where f_k is a factor for a specific combination k of RAC and processing. Values will typically be between 0 and 1, although occasionally the processing factor may also be >1 (*e.g.* drying as applied for grapes and figs).

The user of the model will have to specify processing factors for each food k as defined in the food consumption data base. For this purpose, it is advised to maintain a data base of processing factors, indexed by substance, RAC and processing type (*e.g.* washing, peeling or other processing). Before running the model, it may then be necessary to specify how the necessary processing factors are derived from the data base entries and/or other information. Example: if there are no processing factors known for captan in pears, it may be decided to use the corresponding factors for apples instead.

Often processing effects may be variable, and this may be entered in the Monte Carlo modeling by specifying two values for each processing factor:

1. $f_{k,nom}$: the nominal value, typically some sort of central value from an experimental study
2. $f_{k,upp}$: an upper 95% confidence limit, which typically will be set by an expert (even if statistical information on variability of the factor is available)

A typical data base entry might thus read:

RAC	processing	$f_{k,nom}$	$f_{k,upp}$
apple	washing	0.5	0.7

In the conversion algorithm (see Data Formats document) a check is made on the simultaneous use of Food conversion factors for weight changes between two foods, e.g. a food as eaten and a food as measured (step 3 of the conversion algorithm) and processing factors for concentration changes between these two foods (step 2 of the conversion algorithm). If the same code pair is found in both tables, the processing factor is modified to indicate the change in amount (rather than concentration) of the compound.

In the MC-modeling, processing factors can be used in either of three ways (for each food k to be chosen by the user):

3.1 No processing factor

Just take f_k = 1. This is in most (though not all) cases a worst-case assumption. No data on processing are needed and therefore this route is useful in a first tier approach.

3.2 Fixed processing factors

Use $f_k = f_{k,upp}$. Available information on specific processing effects is used, although still in a cautionary way (in accordance with the precautionary principle). Note that $f_{k,nom}$ values need not to be specified. When both are specified, the highest value will be used; worst case scenario.

3.3 Distribution based processing factors

Sample f_k using a normal distribution. Log or logit transformed values of $f_{k,nom}$ and $f_{k,upp}$ are used to define the first two moments of the normal distribution. Two situations are distinguished depending on the type of transformation.

a) The logarithms of $f_{k,nom}$ and $f_{k,upp}$ are equated to the mean and the 95% one-sided upper confidence limit of a normal distribution. This normal distribution thus is specified by a mean $\ln(f_{k,nom})$ and a standard deviation $\{\ln(f_{k,upp}) - \ln(f_{k,nom})\}/1.645$. Values are drawn from this distribution in the MC-simulations. Processing factors f_k will be nonnegative. Note: $f_{k,upp}$ and $f_{k,nom}$ values equal to 0 are replaced by a low default value (0.01); this is useful computationally to avoid problems with logarithms.

b) The logits of $f_{k,nom}$ and $f_{k,upp}$ are equated to the mean and the 95% one-sided upper confidence limit of a normal distribution. This normal distribution thus is specified by a mean $\text{logit}(f_{k,nom})$ and a standard deviation $\{\text{logit}(f_{k,upp}) - \text{logit}(f_{k,nom})\}/1.645$. Values are drawn from this distribution in the MC-simulations. Processing factors f_k will be between 0 and 1. Note: $f_{k,upp}$ and $f_{k,nom}$ values equal to 0 and 1 are replaced by default values (0.01 and 0.99); this is useful computationally to avoid problems with logits.

The user should keep in mind that, in case of a lognormal distribution, $f_{k,nom}$ defines the median, while $f_{k,upp}$ quantifies skewness. The same holds for the logistic. Usually, a logarithm will be the standard transformation, but for very skew distributions (see Figure 3) occasionally values above 1 are sampled (upper row, 1rst, 3rd and 5th plot). A logit transformation should be considered instead.

To process simultaneously some foods using fixed factors and others distribution based, choose 'processing (distribution based)'. Now, fixed factors f_k are obtained by providing only $f_{k,upp}$ whereas random factors f_k are sampled when both $f_{k,upp}$ and $f_{k,nom}$ are given.

It is not necessary to fill out a complete list of processing factors for all foods. Missing values of $f_{k,nom}$ and $f_{k,upp}$ are, by default, replaced by the value 1.

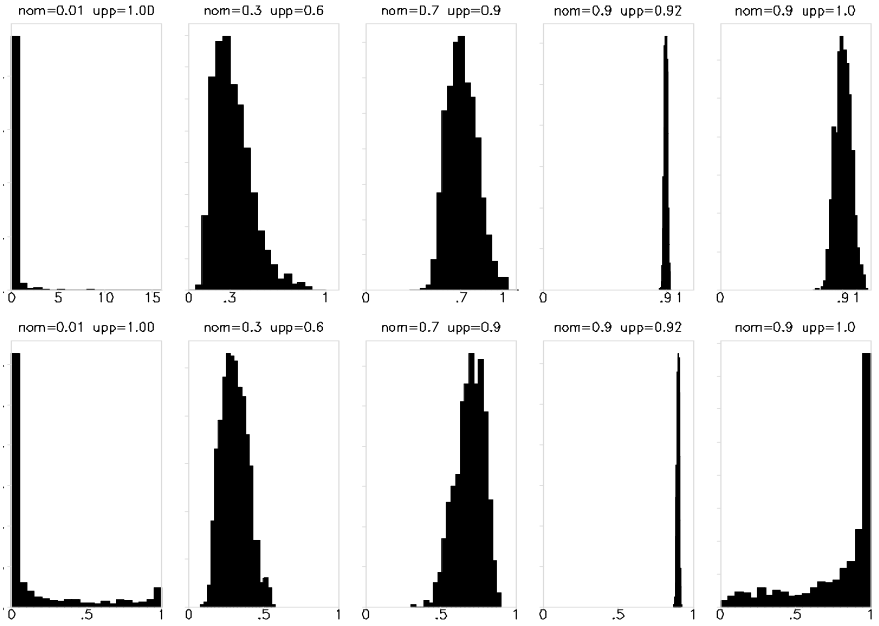


Figure 3: Lognormal (upper row) and logistic (lower row) distributions for various values of $f_{k,nom}$ (= nom) and $f_{k,upp}$ (= upp)

4 Acute risk assessment - Modeling of unit variability

4.1 Variability in composite samples

Variability in concentrations between individual units is a relevant factor in the assessment of short-term dietary intake of substances in food. It is addressed separately because monitoring measurements cm_k are typically made on homogenised composite samples, both in controlled field trials and in food monitoring programs. Such a composite sample for food k is composed of nu_k units with nominal unit weight wu_k each. The weight of a composite sample is therefore $wm_k = nu_k \times wu_k$. This weight is often

larger than a consumer portion, *e.g.* a typical composite sample of 20 sweet peppers weighs 3.2 kg, whereas daily consumer portion weights in the Dutch Food Consumption Survey 1997 ranged from 0.08 g to 458 g.
How should monitoring data be used to estimate the raw food concentration levels c_{ijk} in consumer portions? Although the mean level of cm_k may be a fair estimate of the mean level of c_{ijk}, the variability of cm_k is not appropriate to estimate the variability of c_{ijk}. In smaller portions more extreme values may occur more readily, and thus acute risks may be higher than would follow from a direct use of the composite sample data.
Therefore, the FAO/WHO Geneva Consultation recommended to include a *variability factor* (*v*) in the non-probabilistic calculation of an International Estimate of Short Term Intake (*IESTI*) (FAO/WHO 1997). The *IESTI* has been adopted by the Joint Meeting of FAO and WHO experts on Pesticide Compounds in food in 1999, and was modified in 2000 to reflect that the supply for actual consumption on a given day is likely to be derived from a single lot (JMPR 1999, 2000). In both the original and the modified definition, the variability factor is used in a similar way. The basic idea is that the concentration of a substance for the first unit eaten is multiplied by *v*, whereas this factor is not applied for any remaining part of the daily consumption.

In the original presentation *v* was meant to reflect "*the ratio of a highest concentration in the individual product unit to the corresponding concentration seen in the composite sample*" (FAO/WHO 1997). It was not clearly stated what was meant with "*a highest concentration*". Should this be the maximum concentration found or should it be a high percentile, *e.g.* p95 or p97.5? In practical terms this did not matter too much, because little data were available. Therefore the FAO/WHO Consultation recommended to take *initial* values of *v* equal to "*the number of units in the composite sample as given in Codex sampling protocols*". This will provide a conservative estimate of the concentration of the substance in the first unit, based on the assumption that all of the content of the substance present in the composite sample are present in this single unit. If Codex sampling protocols are used, then the number of units per composite sample is 5 for large crops (unit weights > 250 g) and 10 for medium crops (unit weights 25-250 g). For small crops (< 25 g) a variability factor *v* = 1 was recommended. More recently, it has been proposed to replace the default value 10 with 7. For foods which are processed in large batches, *e.g.* juicing, marmalade/jam, sauce/puree, bulking/blending a variability factor *v* = 1 is proposed. To summarise:

unit weight, *wu*	FAO/WHO default variability factor, *v*
< 25 g	1
25 –250 g	7
> 250 g	5
juicing, marmalade/jam, sauce/puree	1

Table 1: Default variability factors for IESTI calculations

The Consultation specifically recommended to replace these default values with more realistic values obtained from studies on actually measured units. A working group of the International Conference on Pesticide Residues Variability and Acute Dietary Risk Assessment held in York in 1998 suggested to define *v*, for samples taken from controlled trials, as the 97.5th percentile of the unit levels divided by the sample mean (Harris *et al*. 2000), and this is used in the current version of MCRA as the defining relation.

4.2 Deterministic modeling

MCRA offers two possibilities for display of deterministic estimates, in comparison to the probabilistic results:

1) The user can enter an externally calculated value (e.g. from PRIMO see http://www.efsa.europa.eu/en/mrls/mrlteam.htm) for display together with the probabilistic results;
2) the user can ask for calculation of the IESTI.

The IESTI is a deterministic estimate of the short-term intake of a substance on the basis of the assumptions of high daily food consumption per individual and highest concentrations from supervised trials. The IESTI is expressed per kg body weight and has only been defined for single foods.

Calculations of IESTI (according to FAO 2002) recognise four different case (1, 2a, 2b and 3). In cases 1 to 3 the following definitions are used:

LP: Highest large portion reported, calculated as the 97.5th percentile of the distribution of consumed portions on days with positive consumption of the food (kg food/day)

HR: Highest residue (= concentration) in composite sample, mg/kg

bw: Mean body weight, kg; in MCRA values may be input by the user, or weighted means are calculated over individuals with the number of days on which they consumed the food as weights

U: Unit weight of the edible portion, kg.

v: Variability factor – the factor applied to the composite value to estimate the concentration in a high-value unit

MR: Median residue (= concentration) in food, mg/kg

Case 1:
The concentration of the substance in a composite sample reflects the concentration in meal-sized portion of the food (unit weight is below 25 gr).

$$\text{IESTI} = \frac{LP * HR}{bw}$$

Case 2:
The meal sized-portion, such as a single fruit or vegetable unit might have a higher concentration than the composite (whole fruit or vegetable unit weight is above 250 gr). Case 2 is further divided into case 2a and 2b.

Case 2a:
Unit edible weight of raw food is less than large portion weight.

$$\text{IESTI} = \frac{U * HR * v + (LP - U) * HR}{bw}$$

The formula is based on the assumption that the first unit contains concentartions at the HR**v* level and the next one contains concentrations at the HR level, which represents the concentrations in the composite from the same lot as the first one.

Case 2b:
Unit edible weight of raw food exceeds large portion weight.

$$\text{IESTI} = \frac{LP * HR * v}{bw}$$

The formula is based on the assumption that there is only one consumed unit and it contains concentrations at the HR**v* level.

Case 3:
For those processed foods where bulking or blending means that the median represents the likely highest concentration.

$$\text{IESTI} = \frac{LP * MR}{bw}$$

When an acute reference dose is available, the calculated IESTI values are also expressed as a percentage of the ARfD.

4.3 Probabilistic modeling: specifying distributions

How should variability between units be incorporated in probabilistic modeling of acute risks? In probabilistic modeling we generate consumption amounts and concentrations which will be multiplied, summed over foods and divided by body weight to estimate the intake. However, the concentration cm_k will usually be derived from a distribution based on measurements on composite samples. Assume that a batch of food contains N units (N large, for the statistics we assume infinite). The monitoring measurement cm_k is made on a composite sample of nu_k units (for example, $nu_k = 5$). These units are assumed to be representative of the batch. Unit concentrations c_{ijk} are to be simulated for one or more units from this batch that will be part of a consumption portion in the MC-simulation. Basically, there are three possibilities depending on the availability of data:

1. use actual measurement data on individual units;
2. use variability factors or other summary statistics based on measured individual units;
3. use conservative assumptions.

In MCRA methods under categories 2 and 3 are implemented. The first approach has been pioneered in the context of a large UK survey on pesticides in fruit (Hamey 2000).

The following three models are discussed in more detail:

1. **beta model**, requires knowledge of the number of units in a composite sample, and of the variability between units (realistic or conservative estimates);
2. **bernoulli model**, requires only knowledge of the number of units in a composite sample (results are always conservative);
3. **lognormal model**, requires only knowledge of the variability between units (realistic or conservative estimates).

Preferably realistic estimates of unit variability are to be used, either expressed as coefficients of variation *cv* (standard deviation divided by mean) or as variability factors *v* (defined in MCRA as 97.5th percentile divided by mean). However, often such information is not directly available. In such cases it is customary to select high values for the variability factor, either based on collections of variability factors for other substances/foods, or calculated as the theoretical maximum derived from the number of units in a composite sample.

How to translate the concept of conservatism to the probabilistic model? In a non-probabilistic model a higher value of *v* gives a higher *IESTI*, but in a stochastic model a higher variability means more spread around a central value. In general this means that higher values, but also lower values can be generated. In order to retain an overall conservatism it is therefore necessary to replace all simulated values below the monitoring level (cm_k) with cm_k itself.

It is common to use default conservative values, such as the FAO/WHO variability factors in Table 1. However, one should be aware that two entirely different interpretations are possible:

1. The default variability factor may be defined in the same way as a data-based variability factor (v = 97.5th percentile/mean). For example, it may be an expert opinion based on seeing many actual data sets from trials, that a certain value v can be used as a conservative value for other situations (see *e.g.* Table 1 in Harris *et al*. 2000). Then we might use the beta or the lognormal model, censoring these distributions at cm_k to guarantee conservative behaviour. For the beta model additional information on the number of units in a composite sample is needed.
2. Alternatively, one can revert to the original definition and interpret FAO/WHO variability factors as the number of units in the composite sample ($v = nu_k$). In this case, without other information, the only workable model is the bernoulli model.

Table 2 describes the four options when a parametric form for unit variability is specified. Measurements are simulated for a new unit in the batch using a lognormal distribution or for a unit belonging to the composite sample leading to the use of the beta distribution.

	Simulate for new unit in batch (lognormal distribution)	Simulate for unit belonging to composite sample (beta distribution)
Estimates of unit variability are realistic (R)	• no censoring at cm_k • no upper limit to the unit concentration	• no censoring at cm_k • unit values never higher than $nu_k \cdot cm_k$
Estimates of unit variability are conservative (C)	• unit values will be left-censored at cm_k • no upper limit to the unit concentration	• unit values will be left-censored at cm_k • unit values never higher than $nu_k \cdot cm_k$

Table 2: Choices for estimated variability factors. *cm_k* = value of composite sample concentration, *nu_k* = number of units in composite sample.

4.3.1 Beta model

With this model MCRA will generate values for individual unmeasured units of a measured composite sample. If cm_k is the concentration measured (or simulated) for the composite sample in monitoring for food *k*, then the concentration in any unit can be no larger than $c_{max} = nu_k * cm_k$, where nu_k is the number of units in the composite sample. Under the beta model simulated unit values are drawn from a bounded distribution on the interval (0, c_{max}). The parameter for unit variability is specified as a coefficient of variation cv_k or as a variability factor v_k of the unit values in the composite sample.

The standard beta distribution is defined on the interval (0, 1) and is usually characterised by two parameters *a* and *b*, with $a>0, b>0$ (see *e.g.* Mood *et al.* 1974). Alternatively, it can be parameterised by the mean $\mu = a/(a+b)$ and the variance $\sigma^2 = ab(a+b+1)^{-1}(a+b)^{-2}$, or, as applied in MCRA, by the mean μ and the squared coefficient of variation $cv^2 = ba^{-1}(a+b+1)^{-1}$. Note that the coefficient of variation is the same for the unscaled and the scaled distributions.

For the simulated unit values in each iteration of the program we require an expected value cm_k. This scales down to a mean value $\mu = cm_k/c_{max} = 1/nu_k$ in the (standard) beta distribution. From this value for μ and an externally specified value for cv_k the parameters *a* and *b* of the beta distribution are calculated as:

$$a = b(nu_k - 1)^{-1}$$

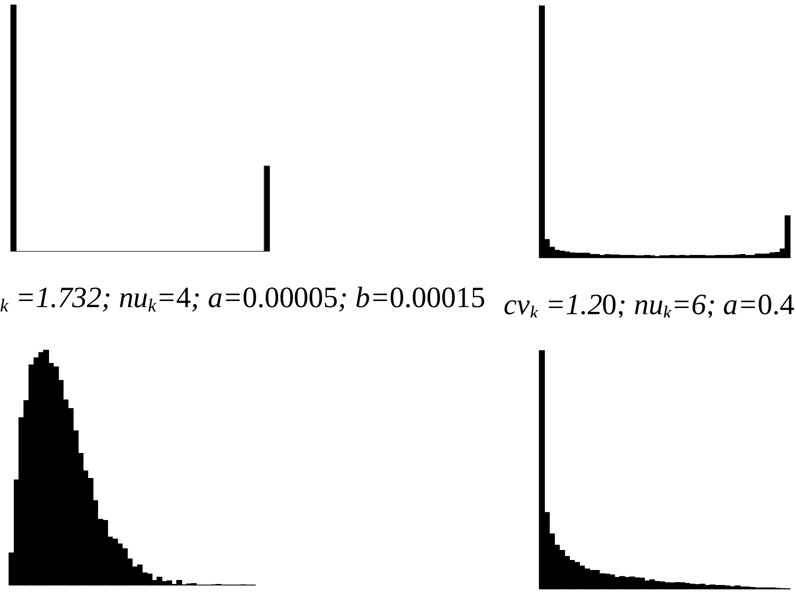
$$b = \frac{(nu_k - 1)(nu_k - 1 - cv_k^2)}{nu_k cv_k^2}$$

From the second formula it can be seen that cv_k should not be larger than $\sqrt{nu_k - 1}$ in order to avoid negative values for *b*.

When the unit variability is specified by a variability factor $v_k = \frac{p97.5_k}{cm_k}$ instead of a coefficient of variation cv_k then MCRA applies a bisection algorithm to find *a* such that the cumulative probability $P[Beta(a,b)] = 0.975$ for $b = a(nu_k - 1)$.

Sampled values from the beta distribution are rescaled by multiplication with c_{max} to unit concentrations c_{ijk} on the interval (0, c_{max}).

In the case that variability has been estimated by a conservative high value, all sampled values lower than cm_k are replaced by cm_k.
In Figure 4, for several values of the coefficient of variation and number of units the beta distribution is shown with estimated parameters a and b. When the parameter for unit variability is high (upper left plot) the ratio of the spikes on the extremes (3:1) represent the 75% probability at $c_{ijk} = cm_k$ and 25% probability at $c_{ijk} = c_{max}$. In the upper right plot, the parameter for unit variability is smaller and some unit values in between the two extremes are sampled. The ratio of the spikes is about 5:1, which is according to the number of units in the composite sample. In the lower left plot, variability is low and unit values are sampled around the monitoring value. In the extreme case, when unit variability is close to zero the monitoring value itself is sampled and a spike occurs (not shown). The lower right plot shows an intermediate situation, moderate to high variability.



$cv_k = 1.732$; $nu_k=4$; $a=0.00005$; $b=0.00015$ $cv_k = 1.20$; $nu_k=6$; $a=0.4$; $b=2$

$cv_k=0.62$; $nu_k=6$; $a=2$; $b=10$ $cv_k = 1.46$; $nu_k=4$; $a=0.1$; $b=0.3$

Figure 4: Standard beta distribution for different values of the coefficient of variation cv_k and number of units nu_k in the composite sample. x axis from 0 to 1.

4.3.2 Lognormal model

With the beta and bernoulli models, MCRA simulates concentrations for units in the composite sample, such that the concentration of an individual unit can never be higher than the monitoring measurement multiplied by the number of units in the composite sample $c_{max} = nu_k * cm_k$.
With the lognormal model for unit variability MCRA simulates concentrations for new units in the batch from which the composite sample was taken. Effectively the number of units in a batch is very large, so in this case there is no practical upper limit to the concentration that can be present.
The lognormal distribution is considered as an appropriate model for many empirical positive concentration distributions. With the lognormal model MCRA assumes a lognormal distribution for unit concentrations. Let this distribution be characterised by μ and σ, which are the mean and standard deviation of the log-transformed concentrations. The unit log-concentrations are drawn from a normal distribution with mean $\mu = \ln(cm_{ik}) - \frac{1}{2}\sigma^2$.
Also for the lognormal model MCRA allows two choices to specify the parameter for the unit variability. The parameter is specified as a coefficient of variation (cv_k) or as a variability factor (v_k). The coefficient of variation cv is turned into the standard deviation σ on the log-transformed scale with:

$$\sigma = \sqrt{\ln(cv^2 + 1)}$$

The conversion of a variability factor into parameters of the lognormal distribution requires an exact definition of what is meant. Here, the variability factor is defined as the 97.5th percentile of the concentration in the individual measurements divided by the corresponding mean concentration seen in the composite sample. A variability factor v is converted into the standard deviation σ as follows:

$$v = \frac{p97.5}{mean} = \frac{e^{\mu + 1.96\sigma}}{e^{\mu + 1/2\sigma^2}} = e^{1.96\sigma - 1/2\sigma^2}$$

with μ and σ representing the mean and standard deviation of the log-transformed concentrations. So

$$\ln(v) = 1.96\sigma - 1/2\sigma^2$$

Solving for σ gives: $\sigma^2 - 2*1.96\sigma + 2\log(v) = 0$, with roots for σ according to:

$$\sigma = 1.96 \pm \sqrt{}(1.96^2 - 2\log(v))$$

The smallest positive root is taken as an estimate for σ .

In the case that variability has been estimated by a conservative high value, all sampled values lower than cm_k are replaced by cm_k.

4.3.3 Bernoulli model

In practice, measurements on individual units to obtain a measure for unit variability are not very common. The bernoulli model is a limiting case of the beta model, which can be used if no information on unit variability is available, but only the number of units in a composite sample is known (see van der Voet *et al*. 2001).

As a worst case approach we may take cv_k as large as possible. When cv_k is equal to the maximum possible value $\sqrt{nu_k - 1}$, the (unstandardised) beta distribution simplifies to a bernoulli distribution with probability $(nu_k - 1)/nu_k$ (or $(v_k\text{-}1)/v_k$) for the value 0 and probability $1/nu_k$ (or $1/v_k$) for the value $c_{max} = nu_k * cm_k$..
In MCRA values 0 are actually replaced by cm_k, to keep all values on the conservative side. For example, with $nu_k = 5$, there will be 80% probability at $c_{ijk} = cm_k$ and 20% probability at $c_{ijk} = c_{max.}$.
When the number of units nu_k in the composite sample is missing, the nominal unit weight wu_k is used to calculate the parameter for unit variability.

4.3.4 Estimation of intake values using the concept of unit variability

- For each iteration i in the MC-simulation, obtain for each food k a simulated intake x_{ik} , and a simulated composite sample concentration cm_{ik} .
- Calculate the number of unit intakes nux_{ik} in x_{ik} (round upwards) and set weights w_{ikl} equal to unit weight wu_k, except for the last partial intake, which has weight $w_{ikl} = x_{ik} - (nux_{ik} - 1)wu_k$.
- For the beta or bernoulli distribution: draw nux_{ik} simulated values bc_{ikl} from a beta or bernoulli distribution. Calculate concentration values as $c_{ikl} = bc_{ikl} * cm_{k,\, max} = bc_{ikl} * cm_k * nu_k$. Sum to obtain the simulated concentration in the consumed portion:

$$c_{ik} = \sum_{l=1}^{nux_{ik}} w_{ikl} c_{ikl} \Big/ x_{ik}$$

- For the lognormal distribution: draw nux_{ik} simulated logconcentration values lc_{ikl} from a normal distribution with mean $\mu = \ln(cm_{ik})$ and standard deviation σ. Back transform and sum to obtain the simulated concentration in the consumed portion:

$$c_{ik} = \sum_{l=1}^{nux_{ik}} w_{ikl} e^{lc_{ikl}} \Big/ x_{ik}$$

5 Chronic risk assessment - Betabinomialnormal model (BBN)

The Betabinomial-Normal (BBN) model for chronic risk assessment is described in de Boer et al. (2009), including its near-dentity to the STEM-II model presented in Slob (2006).

Through the assumed independence of consumption data and concentration values the daily intake of individual i on day *j* can be calculated as the aggregated sum over foods of consumption amount per kg body weight times average concentration. For empirical modeling of concentrations, the average concentration of all available concentration measurements on a food is taken, with non-detect measurements entered as zero, $\frac{1}{2}LOR$ or LOR, or any other fraction of LOR as specified in the input options. For parametric modeling of concentrations, the average concentration per food is calculated as:

model	average concentration for food *k*
empirical	$c_k = \frac{1}{n_k}\sum_{i=1}^{n} x_{ik}$ (nondetect x_{ik} may be replaced by *f* x *LOR*)
mixture of non-detect spike and lognormal	$c_k = (1 - p_k)\mu_k + p_k(f * LOR_k)$
mixture of non-detect spike and truncated lognormal	$c_k = (1 - p_k)\mu_k + p_k(f * LOR_k)$
censored lognormal	$c_k = \mu_k$
censored lognormal with estimated LOR	$c_k = \mu_k$
mixture of zero spike and censored lognormal	$c_k = (1 - p_{0k})\mu_k$

where c_k is the average concentration (on logscale) for food *k*, x_{ik} concentration value, μ_k the estimated mean, p_k the fraction of nondetects, $(1 – p_k)$, the fraction of detects and $p_{0k,}$ the estimated fraction of true zero's (or true nondetects).

5.1 Intake frequency distribution

Let *n* and *npos* be the total number of days per individual (for all individuals equal) and the number of days with a positive intake, respectively. Then *npos* is modelled using a betabinomial distribution with binomial totals *n* and probabilities *p*. The probabilities, *p*, are assumed to follow a beta distribution:

$$\text{f(p)}= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, p^{\alpha-1}(1-p)^{\beta-1}$$

With $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, the probability that *npos* equals *x* can then be written as:

$$P(\underline{npos}=x) = \binom{n}{x}\frac{B(\alpha+x, n+\beta-x)}{B(\alpha,\beta)}, \qquad x = 0, 1 \ldots n$$

This distribution is known as the betabinomial distribution.

The mean and variance of a beta distribution are:

$\alpha/(\alpha+\beta)$
and $\alpha\beta(\alpha+\beta+n)/[(\alpha+\beta)^2(\alpha+\beta+1)]$, respectively.

Re-parameterizing by $\pi = \alpha/(\alpha+\beta)$ and $\varphi = 1/(\alpha+\beta+1)$ is a more stable and interpretable parameterization. It can be shown that the mean and variance of *npos* are equal to $n\pi$ and n $\pi(1-\pi)[1+(n-1)\varphi]$, respectively.

Note that the first part of the variance $n\pi(1-\pi)$ equals the binomial variance; the second part is the so-called overdispersion factor.
Fitting the betabinomial model with maximum likelihood gives estimates $\hat{\pi}$ and $\hat{\varphi}$ for the parameters π and φ. Back-transformation gives the following estimates for α and β :

$$\hat{\alpha} = \hat{\pi}(1-\hat{\varphi})/\hat{\varphi} \text{ and } \hat{\beta} = (1-\hat{\pi})(1-\hat{\varphi})/\hat{\varphi}$$

The distribution of the probability that a individual eats a food with a substance at a certain day is then Beta($\hat{\alpha}$, $\hat{\beta}$).

5.2 Modeling the positive intake amounts

5.2.1 Power or log transformation

First, to achieve a better normality, the positive daily intake amounts are transformed. The user can choose a logarithmic transformation $f(y) = \ln(y)$ (no parameters to be estimated) or a power transformation $f(y) = y^q$ (one parameter to be estimated). In the latter case the optimal power is determined on the grid $\{10, 2, 1, \frac{1}{1.5}, \frac{1}{2}, \frac{1}{2.5}, \frac{1}{3}, \frac{1}{3.5}, \ldots, \frac{1}{100}\}$, with a further refinement grid search around the best fitting value. If a power $\frac{1}{100}$ gives the best fit in this grid search, then the logarithmic transformation is selected (Note that a logarithmic transform corresponds theoretically to $q = 0$). The goodness of fit is determined by minimising the residual sum of squares: $(z(i) - \beta_1 y^q)^2$ of a regression of normal Blom scores on the power-transformed daily intakes. Normal Blom scores are (Tukey 1962):

$$z_{(i)} = \Phi^{-1}\left(\frac{i-\frac{3}{8}}{n+\frac{1}{4}}\right)$$

where *i* is the rank of the n^{th} non-zero daily intake, *n*, the total number of non-zero intakes and $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cumulative distribution function.
In MCRA 7.1 the power transformation, given by y^p, has been replaced by the equivalent Box-Cox transformation. The Box-Cox transformation is a linear function of the power transformation, given by $(y^p-1)/p$, and has a better numerical stability.

5.2.2 Model with between-individual and within-individual variance component

The transformed positive intake amounts are modelled in a ML analysis with random terms *individual* and interaction *individual.day* to estimate the between-individual and within-individual variance component. For a logarithmic transformation :

$$\ln(\underline{y}_{ij}) = \mu + \underline{c}_i + \underline{u}_{ij}$$

and a power transdormation (with power q)

$$\underline{y}_{ij}^{\,q} = \mu + \underline{c}_i + \underline{u}_{ij}$$

where $\underline{c}_i$ and $\underline{u}_{ij}$ are the individual effect and interaction effect, respectively. These effects are assumed to be normally distributed N(0, $\sigma^2_{\ between}$) resp. N(0, $\sigma^2_{\ within}$).

If the positive intake amounts are logarithmically transformed it can be shown that the expectation and variance of the positive intake amount per random consumption day of a random individual are:

$$\mathrm{E}(\underline{y}_{ij}) \quad = \exp(\mu + ½\ \sigma^2_{\ within})$$

$$\mathrm{Var}(\underline{y}_{ij}) = \sigma^2_{\ between}$$

For the power transform the expectation equals:

$$\mathrm{E}(\underline{y}_{ij}) \quad = \quad \mu^{\lambda} + \frac{1}{2}\lambda(\lambda - 1)\mu^{\lambda-2}\sigma^2_{\ within} \quad \text{(Dodd \textit{et al.}, 2006)}$$

5.3 Modeling usual daily intake

5.3.1 Analytical integration

For logarithmic transformed intake amounts, a analytical solution is available (not implemented in the MCRA program).

The usual intake is defined as the intake amount per random intake day (over both intake and non-intake days) of a random individual. To obtain the usual intake the E($\underline{y}$) from 5.2.2 has to be multiplied by the probability π from 5.1 . If π was constant for all individuals the usual intake would have a lognormal distribution with mean ln(π) + μ + ½ $\sigma^2_{\ within}$ and variance $\sigma^2_{\ between}$. But because we have assumed in 5.1 that individuals have different p's coming from a beta distribution, the probability that a individual has a usual intake lower than say an intake limit z equals:

$$\mathrm{P}(py \leq z) = \int_p (\underline{p} = p \wedge \underline{y} \leq \frac{z}{p}) = \int_p (\underline{p} = p \wedge \ln \underline{y} \leq (\ln(\mathrm{z}) - \ln(\mathrm{p})))$$

$$\int_{p=0}^{1} \frac{\Gamma(\hat{\alpha} + \hat{\beta})}{\Gamma(\hat{\alpha})\Gamma(\hat{\beta})} p^{\hat{\alpha}-1}(1-p)^{\hat{\beta}-1}\Phi(\frac{\ln(z) - \ln(p) - \hat{\mu} - ½\,\hat{\sigma}^2_{within}}{\hat{\sigma}_{between}})\,\mathrm{dp}$$

where Φ is the cumulative normal distribution.

5.3.2 Numerical integration

If the positive amounts are transformed by a power transformation the power transformed values can not generally be written in terms of a probability distribution as in 5.3.1 : the distribution of the usual intakes has to be calculated numerically.

However, in the MCRA program for both power and logaritmic transformation, the usual intake distribution is obtained by numerical integration.

The distribution of the usual intakes can be obtained as follows:

1. Draw 1 value from a normal distribution with mean μ and variance $\sigma^2_{between}$
2. Calculate the inverse transformation of the value of Step 1.

   2a. For a logarithmic transformation: $x = \exp(\mu + \sigma_{between}\, ln\, y) + ½\, \sigma^2_{within}$.

   2b. For a power transformation: calculate x by Gauss-Hermite integration. For the Box-Cox transformation the integral can be approximated by means of N-point Gauss-Hermite integration. Gauss-Hermite integration approximates a specific integral as follows

   $$\int_{-\infty}^{\infty} f(x)\exp(-x^2)\ dx \approx \sum_{j=1}^{N} w_j\, f(x_j)$$

   in which w_j and x_j are weights and abscissas for N-point Gauss-Hermite integration, see Abramowitz and Stegun (1972). This results in the following usual intake

   $$x = \frac{1}{\sqrt{\pi}}\sum_{j=1}^{N} w_j\left[1 + p\left(\mu_i + b_i + \sqrt{2}\sigma_{within} x_j\right)\right]^{\frac{1}{p}}$$

   with p the power of the Box-Cox transformation. There can be a small problem with Gauss-Hermite integration. The summation term can not be calculated when the factor between squared brackets is negative and the power p is not an integer. This can happen when $(\mu_i + b_i)$ is small relative to the between day standard error σ_{within}. In that case the corresponding term is set to zero. This is not a flaw in the numerical method but in the statistical model since the model allows negative intakes on the transformed scale which cannot be transformed back to the natural scale.

3. Draw 1 value from the beta distribution
4. Multiply the value of Step 2. with the value of Step 3.

The result is one random draw from the distribution of usual intakes.
Repeat Steps 1 till 4 a great number of times, say 50000.

5.3.3 Extending the models

The intake frequency and transformed intake amount model can be extended to describe the effect of a covariable and/or cofactor. Then, usual intakes are dependent on explanatory variables.

For frequencies:

cofactor:	logit($\underline{\pi}$) = β_{0l}
covariable:	logit($\underline{\pi}$) = $\beta_0 + \beta_1$ f(x_1; df)
both:	logit($\underline{\pi}$) = $\beta_{0l} + \beta_1$ f(x_1; df)
Interaction:	logit($\underline{\pi}$) = $\beta_{0l} + \beta_{1l}$ f(x_1; df)

For amounts

cofactor:	transf($\underline{y}_{ij}$) = $\beta_{0l} + \underline{c}_i + \underline{u}_{ij}$
covariable:	transf($\underline{y}_{ij}$) = $\beta_0 + \beta_1$ f(x_1; df) + $\underline{c}_i + \underline{u}_{ij}$
both:	transf($\underline{y}_{ij}$) = $\beta_{0l} + \beta_1$ f(x_1; df) + $\underline{c}_i + \underline{u}_{ij}$
interaction:	transf($\underline{y}_{ij}$) = $\beta_{0l} + \beta_{1l}$ f(x_1; df) + $\underline{c}_i + \underline{u}_{ij}$

where l=1…L and L is the number of levels of the cofactor, $\underline{y}_{ij}$, the amount, x_1 is the covariable, f is a spline or polynomial function, df the degrees of freedom, $\underline{c}_i$ and $\underline{u}_{ij}$ are the individual effect and interaction effect respectively. These effects are assumed to be normally distributed N(0, $\sigma^2{}_{between}$) resp. N(0, $\sigma^2{}_{within}$). The degree of the function is determined by backward or forward selection.

The usual intake is calculated for the combination of all levels of the cofactor and a specified number of values of the covariable.

6 Chronic Risk Assessment – Logisticnormal-Normal model (LNN0 and LNN)

An alternative to the betabinomial modelling of intake frequencies in BBN model is modelling these frequencies by a logistic normal distribution. In notation, for probability *p*:

$$\text{logit}(p) = \log(p/1\text{-}p) = \mu_i + \underline{c}_i$$

where μ_i represents the person specific fixed effect model and $\underline{c}_i$ represent person specific random effects with estimated variance component $\sigma^2{}_{between}$.

This model is referred to as the LogisticNormal-Normal (LNN) model. The full LNN model includes the estimation of a correlation between intake frequency and intake amount. This is similar to the NCI model described in Tooze et al. (2006).
A simple and computationally less demanding version of the LNN method which does not estimate the correlation between frequency and amount (introduced as ‘LNN’ in MCRA 7.0) has now been renamed LNN0, where the ‘0’ indicates the absence of correlation.

The models are fitted by maximum likelihood, employing Gauss-Hermite integration.

For chronic models amounts are usually transformed before the statistical model is fit. The power transformation, given by y^p, has been replaced by the equivalent Box-Cox transformation. The Box-Cox transformation is a linear function of the power transformation, given by $(y^p\text{-}1)/p$, and has a better numerical stability. Gauss-Hermite integration is used for back-transformation.

A new method to estimate individual usual intakes has been added to BBN and LNN0. The method builds on the proposal of Kipnis et al. (2009), but is modified to ensure that the population mean and variance are better represented. The method is based on shrinkage of the observed individual means (modified BLUP estimates) and shrinkage of the observed intake frequencies. The individual usual intake distribution applies to the population for which the consumption data are representative, and automatically integrates over any covariates present in the model. Individual intakes are not yet available for LNN, and when a covariable is modelled by a spline function of degree higher than 1.

In case of a model with covariates the usual intake is presented in graphs and tables as a function of the covariates (conditional usual intake distributions). In addition, there is also a marginal usual intake distribution which applies to the population for which the consumption data are representative.

When a chronic model is applied to consumption data with just one day per person, MCRA will ask for the input of a variance ratio for the amount model and a dispersion factor for the frequency model. These values can for example be taken from the output of similar exposure assessments of datasets with multiple days per person.

7 Chronic Risk Assessment - Discrete/semi-parametric model (ISUF)

Nusser *et al.* (1996) described how to assess chronic risks for data sets with positive intakes (a small fraction of zero intakes was allowed, but then replaced by a small positive value). The modeling allowed for heterogeneity of variance, *e.g.* the concept that some people are more variable than others with respect to their consumption habits. However, a disadvantage of the method was the restricted use to contaminated foods which were consumed on an almost daily basis, *e.g.* dioxin in fish, meat or diary products. The estimation of usual intake from data sets with a substantial amount of zero intakes became feasible by modeling separately zero intake on part or all of the days via the estimation of intake probabilities as detailed in Nusser *et al.* (1997) and Dodd (1996). In MCRA, a discrete/semi-parametric model is implemented allowing for zero intake and heterogeneity of variance following the basic ideas of Nusser *et al.* (1996, 1997) and Dodd (1996). This implementation of the ISUD model for chronic risk assessment is fully described in de Boer et al. (2009).

Nusser *et al.* (1996, 1997) describe a procedure for the assessment of chronic risks using non-normal dietary intake data. Principally, their method consists of four steps:

1. transforming the daily intake data to approximate normality using a power function or log transformation
2. fitting a grafted polynomial function to the power or log transformed daily intakes. The polynomial provides some flexibility against power transformed components that are still deviating from normality,
3. estimating the parameters of the usual intake distribution in the transformed scale, and
4. estimating the percentiles of the distribution of usual intakes in the original scale.

7.1 Power or log transformation

Daily intakes are calculated as described in section 5 . First, to achieve a better normality, the positive daily intake amounts are transformed. The user can choose a logarithmic transformation $f(y) = \ln(y)$ (no parameters to be estimated) or a power transformation $f(y) = y^q$ (one parameter to be estimated).

7.2 Spline fit

To achieve a better normality, a second transformation (optional) is performed: a spline function $t = g(z)$ is fitted to the logarithmically or power transformed data *t* as a function of the normal Blom scores. The spline function is a grafted polynomial consisting of cubic polynomials between *p* = 3 joint points (knots) and linear functions in the two outer regions. The daily intakes are transformed by interpolating from *t* to $x = g^{-1}(t)$, using the fitted spline function.

After a successful transformation the daily intakes *x* will resemble Blom normal scores and their mean and total variance will therefore be approximately 0 and 1. The normality of the transformed values *x* is checked with the Anderson-Darling test. In the case of a spline transformation, if normality is rejected at the 85% confidence level, then the number of knots *p* is increased and the spline fit is repeated (until a maximum of 22 knots).

7.3 Estimation of the parameters of the usual intake distribution

Variance components for between and within-individual information are fitted to the transformed non-zero daily intakes *x* using the model:

$$x_{ij} = x_i + u_{ij}$$

$$x_i \sim N(\mu, \sigma_B^2); \quad u_{ij} \sim N(0, s_i^2); \quad E(s_i^2) = \sigma_0^2; \quad \text{var}(s_i^2) = \sigma_A^2$$

In this model the total variance of the daily intakes is divided into a between-individual component and a within-individual component. The within-individual variance component can be heterogeneous, that is, it can be different for different individuals. In the model the between-individual variance σ_1^2 and the mean and the variance of the within-individual variance component distribution (σ_B^2 and σ_A^2) are estimated using standard statistical methods (ANOVA). Further, a test statistic *MA4* is calculated to test whether the heterogeneity of variances is significant (see Dodd 1996 for details).

The estimate s_B^2 of the between-individual variance is the basis for the estimation of the distribution of usual intake. The distribution of usual intakes on non-zero intake days in the *x* scale is represented by a set of 400 normal Blom scores (which themselves represent the standard normal distribution) multiplied by s_1: $x_i = s_B z_{(i)}$. The same calculation is applied to user-requested percentiles $z_p = \Phi^{-1}(p)$..

7.4 Back transformation and estimation of usual intake

The 400+ values x_i are back-transformed to the original scale. This is simple if no spline function has been estimated. If a spline function has been used, then it is a rather complicated procedure, because the spline function *g* was developed for daily intakes, not usual intakes. The following steps are made:

1. First the 400+ values x_i are expanded in a set of 9 * 400 values representing the distribution of daily intakes around each of the 400 points;
2. These 9 * 400+ values are back transformed using the functions *g* and *f* , and the sets of 9 values are then recombined (by weighted averaging) into 400 usual intake values y_i ;
3. A spline function g_1 ,especially adapted for usual intakes, is now fitted to the 400 data pairs (x_i, t_i), where $t_i = f(y_i)$;
4. Finally the usual intakes on non-zero intake days are represented by the back-transform using this improved function: $y_i = f(g_1(x_i))$.

The user-requested percentiles y_p are the additional values (*i* > 400) in the 400+ set. The 400 y_i values define the cumulative distribution function by:

$$F(y_i) = \frac{i - \frac{3}{8}}{400 + \frac{1}{4}}.$$

The distribution is adapted in order to account for days with zero intake of individuals (defined here as individuals who have a positive probability of intake on any day, and therefore a non-zero usual intake). This is done by estimating the distribution of individual intake probabilities. This distribution is approximated via a number of classes (*e.g.* 21 or 51, can be selected by the user) arranged by the proportion of days on which there is a positive intake (p_m). Using a binomial distribution for each class, the fraction of individuals in each class ($\theta_m; \quad m = 0,...,M$) is estimated by optimising the fit of the predicted proportions of individuals with 0, 1, 2, ... intake days to the observed proportions. The number of parameters to be estimated is usually higher than the number of possible outcomes for a individual (*e.g.* 3 when there are two days per individual), and therefore a smooth approximation is made using a modified minimum chi-squared estimator. See Dodd (1996) for details. Only the fraction of non-consumers (θ_0) is estimated separately with no restriction to be similar to the other θ_m. It can be noted that the distribution of individual intake probabilities can be better estimated when the number of days per individual in the consumption survey becomes higher. With only 2 days per individual the procedure gives a rather artificial distribution, often with an estimated θ_0 of zero This step can be time-consuming. Therefore, the number of iterations in the estimation procedure can be limited by the user. In our experience it is not generally necessary to use 50,000 iterations as in Dodd (1996).

The estimated distribution of individual intake probabilities ($\hat{\theta}_0,...,\hat{\theta}_M$) is used to transform the distribution of usual intake on non-zero intake days (F_y) to the distribution of usual intake for individuals (F_C) and finally to the distribution of usual intake for the entire population (F_U). These transformations are based on the relation:

$$F_U(u) = \theta_0 + \sum_{m=1}^{M} \theta_m F_y(u/p_m)$$

which basically says that to obtain a certain level of usual intake *u* we should consider a different level (*u*/p_m) for the class of individuals which consume only on a fraction p_m of days. See Dodd (1996) for details of the computational procedure. Linear interpolation based on the 400 values of the F_y distribution is then used to compute representations of the cumulative distribution functions for individuals only and the entire population.

8 Chronic Risk Assessment - Observed individual means (OIM)

The usual intake distribution for a population is estimated with the empirical distribution of within-individual means. Each mean is the average of all single-day intakes (see 5) for an individual. The mean value for an individual still contains a considerable amount of within-individual variation. As a consequence, the distribution of within-individual means has larger variance than the true usual intake distribution and estimates using the OIM-method are biased, leading to a too high estimate of the fraction of the population with a usual intake above some standard.

9 Acute risk assessment and the BBN model

An acute risk assessment may be followed by an analysis where the acute intake distribution is related to a covariable and/or cofactor. Through MC-sampling, a large number of intakes is generated by combining randomly chosen consumption patterns of individuals *i* on day *j* with randomly chosen concentrations in the consumed foods. The replicates generated for individual day *ij* are further indexed by *k* to represent differences due to concentration variability. We ignore the finiteness of the concentration data, that is, we ignore the identity of the chosen concentration values in the original concentration dataset.

9.1 Intake frequency model

Let n_i and $npos_i$ be the total number of simulated intakes per individual, and the number of simulated positive intakes, respectively. Then $npos_i$ is modelled as a function of *e.g.* age (and/or other individual characteristics), using a betabinomial distribution with binomial totals n_i and overdispersion parameter ϕ (independent of age). The fitted binomial probabilities are $\hat{\pi}_x = f(x_i)$, where x_i is the age of individual *i*, and the estimated overdispersion parameter is $\hat{\phi}$.

9.2 Intake amount model

For the positive intakes, consider power of logarithmically transformed values y_{ijk}. (see 5.2.1) Average over replicates to obtain individual day averages $y_{ij.}$. These values are modelled in a ML analysis with random terms individual and individual.day as a function of age (and/or other individual characteristics), with the number of values per individual day (n_{ij}) as weights w_{ij} to correct for differences in the precision at the individual day stratum. The fitted values from the model are $\hat{\mu}_x = f(x_i)$, where x_i is the age of individual *i*

9.3 Estimating the acute risk variability of positive intake amounts

Correct the full set of simulated positive intakes by $y'_{ijk} = y_{ijk} - \hat{\mu}_{x(i)}$. Estimate the variance $\sigma^2_{y'}$ of y'_{ijk} . We denote the estimated variance as $\hat{\sigma}^2_{y'}$. Now for each selected age *x* the transformed positive intake distribution is modelled as normal with mean $\hat{\mu}_x = f(x)$ and variance $\hat{\sigma}^2_{y'}$.

9.4 Estimating the acute intake distribution

Acute intake distributions dependent on a covariate are obtained by numerical integration. For each combination of levels of the covariable and cofactor, intake frequency values and transformed intake amounts are simulated and multiplied. This results in a number of distributions each one representing the acute intake distribution corresponding to a specific combination of levels of the covariates.

10 Market shares of brands and brand loyalty

Different brands of a food product may differ in levels of chemical substances or contaminants. For example, brands of potato crisps may have different levels of acrylamide due to differences in the sugar content of used potato varieties or differences in baking procedure (temperature, baking time). If both consumption data and concentration data are measured at the brand level, this presents no special difficulties for exposure assessment. Each brand is then just treated as a separate product. However, in practice brand information is often available for the concentration data, but not for the consumption data. In such cases it is necessary to have additional information on the consumption of the different brands.

For a certain type of product, market share data may be available specifying the percentages of each brand. Ideally these percentages are weight percentages, but sales percentages may be used as a proxy.

10.1 Acute health effects

For acute health effects market shares are all that is needed to adapt a probabilistic exposure assessment. For each simulated consumption a brand can be selected with a probability proportional to the market share, and then a concentration value can be sampled from the distribution of concentrations specific for that brand.

10.2 Chronic health effects

In the case of chronic health effects we need additional information. It now becomes important to know if individuals always consume the same brand, or that they consume different brands, thus effectively averaging the concentrations of the different brands in their long-term food intake. The tendency to repurchase the same brand has become known as *brand loyalty*.

There are two main approaches for modeling brand loyalty, known as the stochastic and deterministic approach (Odin *et al*. 2001). Whereas the stochastic approach just tries to give a satisfying description of observed brand loyalty behaviour, the deterministic approach tries to analyse the attitude of individuals towards brand selection in terms of a limited number of explanatory factors. In the context of dietary risk assessment it is typically the stochastic approach which is more useful.

There is a simple stochastic model which has turned out to be extremely useful in analysing buying behaviour. This is the socalled *Dirichlet* model, first given in a comprehensive form by Goodhardt *et al.* (1984). The surprising feature of this model is that it contains only one parameter for brand loyalty, implying that brand loyalty varies little, or relatively little, between competitive brands (Ehrenberg *et al*. 2004). Although this may seem a too simple representation at first, it has been found to give a close description of actual buyer behaviour in most cases of a systematic check across 34 products categories (Uncles *et al.* 1994).

10.3 The Dirichlet model adapted for probabilistic exposure assessment

In a probabilistic model for chronic exposure asessment when brands are known for concentration data, but not for consumption data, we need the following information:

1. the distribution of consumption by individuals on multiple days;
2. for each brand: the distribution of concentrations in that brand of product;
3. market shares of all brands of a product, and a brand loyalty factor L.

Typically 1 and 2 will be in the form of empirical datasets, for example resulting from food consumption surveys and monitoring programmes, respectively. Alternatively, we can specify parametric distributions, with parameters that are fitted to data or just specified based on prior knowledge or assumptions in what-if scenarios.
Technically the Dirichlet model for brand choice needs *nbrand* parameters α_i (which should be positive real numbers). The average brand choice probability for each brand is α_i/S, where $S = \sum \alpha_i$. By definition, the market shares m_i should be proportional to the brand choice probabilities, and thus to the parameters α_i. This means that S, the sum of the alphas, is the only additional parameter that should be specified, and indeed this is the parameter that determines brand loyalty. S=0 corresponds to absolute brand loyalty, and brand loyalty decreases with increasing S. We define $L = (1+S)^{-1}$ as an interpretable brand loyalty parameter, where now $L = 0$ and $L = 1$ correspond to the situations of no brand loyalty and absolute brand loyalty, respectively.

Given empirical or parametric distributions of consumption and concentration values, the algorithm for chronic exposure assessment now operates as follows:

1. collect consumptions for a large number of *n* of individuals,
2. in case of market shares: simulate *n* selection probabilities from the Dirichlet distribution,
3. estimate intake y_{ijk} for individual *i* on day *j* for food *k* as the weighted sum of the average concentration for *B* brands times consumption x_{ijk} and standardize for body weight. In notation: $y_{ijk} = \frac{\sum_{b=1}^{B} x_{ijk} c_{kb} bcp_{ikb}}{w_i}$, where weight bcp_{ikb} is the brand choice probability *b* for food *k* of individual *i*, c_{kb} is the average concentration for brand *b* of food *k*. Note that $\sum_{b=1}^{B} bcp_{ikb}$ sums to 1 for each individual *i*,
4. aggregate intakes over the number of foods *p*,
5. proceed as usual

11 Uncertainty analysis

In probabilistic risk assessment of dietary intake we use distributions which describe the *variability* in consumption within a given population of individuals and the *variability* of the occurrence and level of substances in the consumed foods. However, these calculations do not consider the amount of *uncertainty* that is due to the limited size of the underlying datasets. Typically, in a large number of simulations very many different combinations of consumption and concentrations are made. This leads to a smooth distribution of simulated intakes, and the impression of a very precise estimation of intake percentiles or other quantities of interest. It is essential to realise that the accuracy of the inference depends on the accuracy of the basic data.

When doing an uncertainty analysis in MCRA a number of iterations is chosen, and in each iteration new inputs are resampled for a complete Monte Carlo analysis:

1. Datasets (concentration data, individual data) are resampled from the original database (bootstrap methodology)
2. Parametric inputs, such as portion size uncertainty and processing factors and their variabilities are resampled from parametric distributions.

Even after consideration of sources of uncertainty that can be quantified, many uncertainties remain that cannot be quantified easily. EFSA (2006) has indicated a general way to handle such qualitative uncertainties. MCRA contains a tool developed by Helen Owen (FERA, UK) to build a table addressing qualtitative uncertainties and add this table to the output of an MCRA run.

11.1 Resampling datasets

A computer-based instrument to assess the reliability of outcomes is the *bootstrap* (Efron 1979, Efron & Tibshirani 1993). In its most simple, non-parametric form, the bootstrap algorithm resamples a dataset of *n* observations to obtain a *bootstrap sample* or *resampled set* of again *n* observations (sampling with replacement, that is: each observation has a probability of 1/*n* to be selected at any position in the new resampled set). By repeating this process *B* times, one can obtain *B* resampled sets, which may be considered as alternative data sets that might have been obtained during sampling from the population of interest. Any statistic that can be calculated from the original dataset (*e.g.* the mean, the standard deviation, the 95th percentile, etc.) can also be calculated from each of the *B* resampled sets. This generates a *uncertainty distribution* for the statistic under consideration. The uncertainty distribution characterises the uncertainty of the inference due to the sampling uncertainty of the original dataset: it shows which statistics could have been obtained if random sampling from the population would have generated another sample than the one actually observed.

In MCRA, two type of data are combined: individual consumption data and concentration data. It makes sense to apply resampling to both type of data separately, in order to characterise the uncertainty in the final intake. In MCRA the uncertainty algorithm (when selected) is applied to:

1. the multivariate consumption patterns and associated body weights: actually the data set of individuals is resampled, and all individual information (consumption patterns for all consumption days, body weight, and age) is coupled to the selected individual.
2. the univariate concentration data sets: these are resampled independently for all foods. In principle, the uncertainty algorithm is applied to the dataset consisting of both non-detects and positive values; in practice, for a dataset with n_0 non-detects and n_1 positive values, the number of positive values in a resampled set is obtained as a draw from a binomial distribution with parameter $n_1/(n_0+n_1)$ and binomial total n_0+n_1. Then, this number of values is selected randomnly from the set of n_1 positive values.

In MCRA the resulting uncertainty distribution of percentiles of the intake distribution is summarised by specifying empirical 2.5th , 25th, 75th and 97.5th percentiles. The outer percentiles constitute a central 95% confidence interval for the variability percentiles. However, for this it is necessary that the number of resampled sets *B* is high enough. The number of resampled sets should be chosen depending on the confidence level wanted for the uncertainty interval. Typically 500-2000 resampled sets will be reasonable for a 95 % confidence interval (Efron & Tibshirani 1993, pp. 14-15, 275).

The same uncertainty algorithm can also be applied to deterministic estimates which are calculated from data sets. For example the maximum concentration found in a resampled set will be different, if the actual maximum value in the original dataset has *not* been selected. Also data-based estimates of large portion and average body weight will vary.

11.2 Resampling parametric distributions, processing

Processing effects are modelled either by a fixed processing factor, or by a lognormal or logistic-normal distribution (depending on the distribution type set in table Processingfactor).
In the former case (fixed factor) the uncertainty distribution is lognormal or logistic-normal with the same mean μ as the fixed value, and with a standard deviation σ_{unc} which is calculated from the specified central; value (procnom) and an estimate of p95 of the uncertainty distribution (procnomuncupp).
The calculation for the logistic-normal distribution (disttype 1):

$$\sigma_{\mathrm{unc}} = \{\mathrm{logit}(\mathrm{procnomuncupp}) - \mathrm{logit}(\mathrm{procnom})\}/1.645,$$

and for the lognormal distribution (disttype 2).

$$\sigma_{\mathrm{unc}} = \{\ln(\mathrm{procnomuncupp}) - \ln(\mathrm{procnom})\}/1.645,$$

Values lower than 0.01 or higher than 0.99 (disttype 1 only) are replaced by default values (0.01 and 0.99); this is useful computationally to avoid problems. In each iteration of the uncertainty analysis a new value is drawn from this distribution to be used as a fixed factor in the Monte Carlo calculation. In the case of a processing factor distribution (describing the variability of processing factors) two uncertainties can be specified. First, the uncertainty about the central value μ can be specified as before using a parameter procnomuncupp. Secondly, the uncertainty about the variability standard deviation σ_{var} can be specified by the number of degrees of freedom *df* of a modified chi-square distribution which is used to generate new values of σ_{var}. Setting *df* very high means litte uncertainty, and σ_{var} will be almost equal in all iterations of the uncertainty analysis. Setting *df* close to 0 means a large uncertainty, and very different values of σ_{var} will be obtained in the iterations of the uncertainty analysis.

11.3 Portion size uncertainty

In the context of the European Food Consumption Validation Project (EFCOVAL) the MCRA model for uncertainty has been adapted specifically to the six quantification methods of EPIC-SOFT (Table 3). Using EPIC-SOFT for 24-hour recall consumptions are quantified using portion size and amounts of portions consumed. Although individual consumption data are expressed in grams per day, the primary data may be associated with uncertainty in portion size and amount or number of portions consumed. So, the primary data are unitweights (e.g. the weight of a portion shown on a photo, or the weight of a standard household measure) and amounts of units (e.g. the number of shown portions or the number of cups), the multiplication of both values is the amount consumed in grams. The corresponding portion size uncertainty is primarily connected with unitweights and amounts.

Method	Unitweight (*uw*)	Amount (*a*)
Photographs (P)	Standard portion in grams (Photo 1 of broccoli is 78 g)	Proportion or multiple of standard portion (1 times photo 1 of broccoli)
Household measures (H)	Standard portion in grams (a glass of tea is 150 g)	Proportion or multiple of standard portion (2 glasses of tea)
Standard units (U)	Standard portion in grams (a can of corn is 285 g)	Proportion or multiple of standard portion (1/2 a can of corn)
Standard portion (S)	Standard portion in grams (onion along with fries weighs 10 g)	1
Gram/volume (G)	1	Amount in grams (75 g of potato salad)
Unknown (?)	1	Amount in grams (Salad dressing weighs 15 grams)

Table 3: Overview of EPIC-SOFT quantification methods, with examples in brackets

Three methods (P, H and U) use both unitweights and amounts, one method (S) uses only unitweights, and two methods (G and ?) use only amounts. The difference between unitweight and amount is as follows: unitweights (in grams) are unique for a specific “food item – quantification method”-combination, but the same for all individuals in the survey, whereas amounts are potentially different for each food item on each eating occasion for each day of an individual. Amounts are in grams (methods G and ?) or in number of units (methods P, H, and U).

For an uncertainty analysis of the usual intake assessment of foods and nutrients three sources of uncertainty are modelled:

1. sampling uncertainty of the set of individuals interviewed on their consumption,
2. uncertainty in *uw* (for EPIC-SOFT quantification methods P, H, U and S)
3. uncertainty in *a* (for EPIC-SOFT quantification methods G, P, H, U and ?).

For quantification methods P, H and U the uncertainty in uw as well as the uncertainty in a needs to be specified, for quantification methods G and ? the uncertainty in a needs to be specified, and for method S the uncertainty in uw needs to be specified (Table 3). The uncertainty *cv* specifications were obtained using limited expert opinion to provide estimated upper values for a and uw, and equating these to the p97.5 of the (log)normal uncertainty distribution (the best estimates are interpreted as the mean m).

More details of the approach to portion size uncertainty implemente din MCRA are described in Souverein et al. (2011).

11.4 Measurement uncertainty

An option has been added to correct for measurement uncertainty. Concentration data are variable because of sampling error, but also because of analytical error. The amount of analytical error is commonly quantified as measurement uncertainty. If the measurement uncertainty of the concentration data is known it can be subtracted from the total variance in the data (to a minimum of 0). For measurement uncertainty MCRA will suggest a coefficient of variation (CV) based on the Horwitz equation and the mean level of the compound over all available concentration data. When analytical error is reported by the analytical laboratory, then this information could be used (CV, as %).

12 Simulated intake data

MCRA can also be used to simulate intake data for one or two foods based on an extended version of the LNN model. In addition to the simulated data the true usual intakes can be stored.

There are two sources of variation with respect to intake: between individuals and between days within individuals. These sources are handled separately in the simulation model. The description below is in terms of logarithmic or Box-Cox transformed amounts. Only at the very end the simulated amounts are back-transformed to the original scale. The simulation model is discussed below for two foods. This model is similar to that in Slob et al. (2010), but here it is described in such a way that extensions to multiple foods follows readily.

Variation between individuals

First consider the variation between individuals. The frequency with which an individual consumes a food is given by the Logistic-Normal distribution, i.e. for foods 1 and 2 the individual-based frequencies P_1 and P_2 are given by

$$\text{logit}(P_1) = F_1 \qquad \text{with } F_1 \sim \text{Normal}\left(\eta_1, \sigma_{1f}^2\right)$$
$$\text{logit}(P_2) = F_2 \qquad \text{with } F_2 \sim \text{Normal}\left(\eta_2, \sigma_{2f}^2\right)$$

The individual-based (transformed) amounts A_1 and A_2 are given by normal distributions:

$$\mathrm{A}_1 \sim \text{Normal}(\mu_1, \sigma_{1a}^2)$$
$$\mathrm{A}_2 \sim \text{Normal}(\mu_2, \sigma_{2a}^2)$$

Instead of four independent normal distributions, a 4-dimensional multivariate normal distribution can be specified for (F_1, F_2, A_1, A_2) with mean vector $M_I = (\eta_1, \eta_2, \mu_1, \mu_2)$ and variance-covariance matrix

$$V_I = \begin{pmatrix} \sigma_{1f}^2 & & & \\ \rho_{1a1f}\sigma_{1a}\sigma_{1f} & \sigma_{1a}^2 & & \\ \rho_{2f1f}\sigma_{2f}\sigma_{1f} & \rho_{2f1a}\sigma_{2f}\sigma_{1a} & \sigma_{2f}^2 & \\ \rho_{2a1f}\sigma_{2a}\sigma_{1f} & \rho_{2a1a}\sigma_{2a}\sigma_{1a} & \rho_{2a2f}\sigma_{2a}\sigma_{2f} & \sigma_{2a}^2 \end{pmatrix}$$

Note that in this notation the first subscript (1 or 2) denotes the food while the second subscript denotes whether the between persons variability relates to frequency (f) or to amount (a). The 6 correlations in the matrix above relate to between-persons variability. These can be grouped as follows

- The correlation between the frequency and amount of the same food (ρ_{1a1f} and ρ_{2a2f})
- The correlation between the frequency of food 1 and the frequency of food 2 (ρ_{1f2f})
- The correlation between the amount of food 1 and the amount of food 2 (ρ_{2a1a})
- Mixed correlations between the frequency of one food and the amount of the other food (ρ_{2f1a} and ρ_{2a1f}).

Variation between days within individuals

A random draw from the 4-dimensional normal distribution given above results in frequencies $p_1$and p_2 and amounts a_1 and a_2 for an individual. The day-to-day variation for this individual can be modeled in a similar fashion. Following Slob et al (2010) the frequencies $p_1$and p_2 are first transformed to the real line by means of $z_1 = \Phi^{-1}(p_1)$ and $z_2 = \Phi^{-1}(p_2)$, in which Φ^{-1} is the inverse of the normal distribution function. The frequencies and amounts on a particular day for this individual can then be modeled by a 4-dimensional multivariate normal distribution for (Z_1, W_1, Z_2, W_2) with mean vector $M_D = (0, 0, 0, 0)$ and variance covariance matrix

$$\mathrm{V_D} = \begin{pmatrix} 1 & & & \\ \xi_{1z1w}\sigma_{1w} & \sigma_{1w}^2 & & \\ \xi_{1z2z} & \xi_{1w2z}\sigma_{1w} & 1 & \\ \xi_{1z2w}\sigma_{2w} & \xi_{1w2w}\sigma_{1w}\sigma_{2w} & \xi_{2z2w}\sigma_{2w} & \sigma_{2w}^2 \end{pmatrix}$$

A random draw (Z_1, W_1, Z_2, W_2) for a particular day results in the following simulated intake for food 1 on that day, and similarly for food 2. The first food is consumed when the random draw Z_1 is smaller than z_1. When the food is consumed, the transformed amount consumed is given by $a_1 + W_1$. Finally this amount is transformed to the original scale by using the inverse of the logarithmic or Box-Cox transformation, i.e. $y_1 = \mathrm{g}^{-1}(a_1 + W_1)$in which $\mathrm{g}^{-1}()$ denotes the inverse transformation.

The back transformation $z_1 = \Phi^{-1}(p_1)$ in combination with the marginal normal distribution for Z_1 ensures that the long-term frequency of consuming the first food equals p_1. For the frequencies the model uses two transformations: the logistic transformation for the individual based frequencies and the probit, i.e. Φ^{-1}, for the day-to-day transformation. An alternative would be to use the probit or logit transformation for both.

Generalization to multiple foods

The approach given above is easily extended to multiple foods. The number of parameters increases quickly with the number of foods. With *n* foods the parameters can be grouped as follows

- 2*n* mean parameters for individuals, i.e. η_i and μ_i for each food.
- 2*n* variance parameters for individuals, i.e. σ_{fi}^2 and σ_{ai}^2 for each food.
- *n*(2*n*-1) correlation parameters for individuals, i.e. ρ.
- *n* variance parameters for days, i.e. σ_{wi}^2 for each food.
- *n*(2*n*-1) correlation parameters for day, i.e. ξ.

Restrictions on the variance-covariance matrices

Both variance-covariance matrices V_I and V_D must be positive definite. This implies that the correlations cannot be chosen independently. This can be seen as follows. Suppose that the correlation between variables A and B equals 0.9 and the correlation between A and C also equals 0.9. It is clear that the variables B and C are then also positively correlated. It can be shown that the correlation between B and C must then be larger than 0.62, and this is also the case when the two correlations are

-0.9. To explore this a bit further in the current context, suppose that we have 2 foods and we wish to set the mixed individual correlations (ρ_{2f1a} and ρ_{2a1f}) to zero. Further suppose that the two correlations (ρ_{1a1f} and ρ_{2a2f}) between frequency and amount of the same food have the same value, say *m*, and the two correlations between frequencies and between amounts (ρ_{2f1f} and ρ_{2a1a}) have the same value as well, say *r*. We then have the following correlation matrix, with variances set to 1:

$$\mathrm{V_D} = \begin{pmatrix} 1 & & & \\ \mathrm{m} & 1 & & \\ \mathrm{r} & 0 & 1 & \\ 0 & \mathrm{r} & \mathrm{m} & 1 \end{pmatrix}$$

It can be shown that when *m* = 0.8 the value of *r* must be in the interval (-0.2, 0.2), and the same holds when *m* = -0.8. Generally, for this correlation matrix, *r* must be in the interval (-(1-*m*), 1-*m*). With the additional constraint that *r* = *m*, the correlations must be in the interval (-0.5, 0.5). When the six correlations are all set to the same value, this value must be in the interval (-1/3, 1).

Usual intake

The usual intake of an individual is defined as the long-term intake which is obtained by integrating out the day-to-day variability. In mathematical terms it is the expectation of the intake on a single day (for a given individual). In the two foods case suppose that (f_1, f_2, a_1, a_2) is a realization of (F_1, F_2, A_1, A_2) for an individual. The intake of food i on a single day is defined as

$$y_i = 0 \quad \text{with probability } (1 - p_i)$$
$$y_i = g^{-1}(a_i + W_i) \qquad \text{with probability } p_i \text{ and}$$

$W_i \sim \mathrm{Normal}(0, \sigma_{wi}^2)$

The long-term intake for this person, summed over foods, then equals

$$\mathrm{E(y_1 + y_2) = \ Ey_1 + Ey_2 = logit^{-1}(f_1)E_W}g^{-1}\mathrm{(a_1 + W_1) + logit^{-1}(f_2)E_W}g^{-1}\mathrm{(a_2 + W_2)}$$

and this result readily extends to more than 2 foods. The day-to-day correlations ξ are thus of no importance for the usual intake. For the log-transform the expectation $\mathrm{E}_w g^{-1}(a_i + W_\mathrm{i})$ is given by $\exp(a_i + \sigma_{iw}^2/2)$ which follows immediately from the expectation of the log-normal distribution. For the Box-Cox transformation the expectation can be approximated by Gauss-Hermite integration. The usual intake distribution can approximated by simulating many individuals and calculation the usual intake for these individuals. This gives a large sample from the usual intake distribution which can be used to calculate quantiles.

Below, expressions are given for the mean and variance of the usual intake distribution (i.e. integrated over individuals) of a single food. Although the mean and variance of the usual intake distribution have little practical relevance, they can be used as a first check of the simulation results. The mean and variance can be calculated directly when the following two conditions are met:

- there is no correlation between the frequency and the amount, i.e. ρ_{1a1f} and ξ_{1z1w} both equal zero. This implies that the frequency and amount are independent,
- the log transform is used for the intake.

In that case it follows from the expectation of the log-normal distribution that for an individual with a realization a of *A* the true usual intake equals $\exp(a + \sigma_w^2/2)$. Since *A* itself follows a normal distribution with mean μ and variance σ_a^2 it is clear that for the usual intake distribution:

$$\mathrm{log(Y) \sim Normal(\mu + \sigma_w^2/2\,, \sigma_a^2)}$$

The expectation and variance of y are then given by

$$\mu_\mathrm{Y} = \exp(\mu + \sigma_\mathrm{a}^2/2 + \sigma_\mathrm{w}^2/2)$$
$$\sigma_\mathrm{Y}^2 = [\exp(\sigma_\mathrm{a}^2) - 1]\ \mu_\mathrm{y}^2$$

The expectation and variance of the frequency P can be readily obtained by numerical integration giving μ_P and σ_P^2, i.e.

$$\mu_P = \int [1 + \exp(-x)]^{-1}\, \varphi(x;\, \eta;\, \sigma_f^2)dx$$
$$\sigma_P^2 = \int \{[1 + \exp(-x)]^{-1} - \mu_P\}^2\, \varphi(x;\, \eta;\, \sigma_f^2)dx$$

in which $\varphi()$ is the normal density function with mean η and variance σ_f^2.
The expectation and variance of the usual intake distribution are then given by

$$E(P \cdot Y) = E(P) \cdot E(Y) = \mu_P \mu_Y$$
$$Var(P \cdot Y) = E(P^2 \cdot Y^2) - E^2(P \cdot Y) = (\sigma_P^2 + \mu_P^2)(\sigma_Y^2 + \mu_Y^2) - \mu_P^2\, \mu_Y^2$$

If necessary higher order moments can be calculated in a similar way.

13 About MCRA

MCRA is a result of an ongoing co-operation between RIKILT/RIVM and Biometris since 1998 (where RIKILT Intitute of Food Safety was the home of the current RIVM activities until 2009). RIVM co-ordinates the Dutch KAP programme (Quality of Agricultural Products) where results of monitoring programs for chemical substances in food are gathered in a national database. RIVM also has a recipe database to link food codes from the Dutch food consumption table to primary agricultural products. Biometris contributes statistical models and programs for quantitative risk analysis.
Since 2005, the program was extended in collaboration with RIVM to include models similar to those available in the STEM, SPADE and NCI softwares.

MCRA is written in **Microsoft Visual C# .NET 2008**. MCRA is internet-based and can be used by registered users at http://mcra.rivm.nl. It consists of a basic program to do the computations and of additional database selection possibilities implemented in ASP.NET. A R-(D)COM interface is used to connect the application with R, which is running in the background for statistical analyses and graphics (http://cran.r-project.org).

An earlier version of the MCRA program, as well as an implementation of the Monte Carlo method in @Risk (1996), have been described in van der Voet *et al*. (1999), further elaboration was given in de Boer & van der Voet (2000, 2001, 2006), de Boer *et al.* (2009) and van der Voet *et al*. (2001).

This manual covers the current release 7.1 (release 7 version 1). Major updates of the program, encompassing new or improved facilities will be released with an increased release number and a new manual.

Find more information about the current MCRA release in:
MCRA 7.1 Reference Manual
MCRA 7.1 Overview
MCRA 7.1 Data Formats
MCRA 7 Examples
MCRA 7.1 On Line Help

14 References

- @Risk (1996). Advanced risk analysis for spreadsheets, Windows version. Pallisade Corporation, Newfield, NY, USA.
- Bestfit.(1997). Probability distribution fitting for Windows. Pallisade Corporation, Newfield, NY, USA.
- Blom, G (1958). Statistical estimates and transformed beta-variables. Wiley, New York.
- Boon, P.E., Bonthuis, M., van der Voet, H. and van Klaveren, J.D. (2011). Comparison of different exposure assessment methods to estimate the long-term dietary exposure to dioxins and ochratoxin A. *Food and Chemical Toxicology*, 49: 1979-1988.
- Boon, P.E., Svensson, K., Moussavian, S., van der Voet, H., Petersen, A., Ruprich, J., Debegnach, F., de Boer, W. J., van Donkersgoed, G., Brera, C., van Klaveren, J. D. and Busk, L. (2009). Probabilistic acute dietary exposure assessments to captan and tolylfluanid using several European food consumption and pesticide concentration databases. *Food and Chemical Toxicology*. 47: 2890-2898.
- Boon, PE, van der Voet, H & van Klaveren, JD (2003). Validation of a probabilistic model of dietary exposure to selected pesticides in Dutch infants, *Food Additives and Contaminants*, 20, Suppl. 1: S36-S49.
- Crossley, SJ (2000). Joint FAO/WHO Geneva consultation – acute dietary intake methodology. *Food Additives and Contaminants*, 17: 557-562.
- David, HA (1970). Order statistics. John Wiley & Sons, New York.
- de Boer, WJ, van der Voet, H, PE Boon, G van Donkersgoed & JD van Klaveren (2004). MCRA, a web-based program for Monte Carlo Risk Assessment, Release 3, User Manual. Report March 2004. Biometris and RIKILT, Wageningen University and Research Centre, Wageningen.
- de Boer, WJ, van der Voet, H (2007). MCRA Rel 6, a web-based program for Monte Carlo Risk Assessment, User Manual. Report Nov 2006. Biometris, Wageningen University and Research Centre, Wageningen.
- de Boer, W.J., Voet van der, H., Bokkers B.G.H., Bakker, M.I., Boon, P.E. (2009). Comparison of two models for the estimation of usual intake adressing zero consumption and non-normality. Food Additives and Contaminants. Part A, 26:11,1433 – 1449.
- de Boer, WJ & van der Voet, H (2000). Dietary risk assessment concerning acute exposure to residues and contaminants using summary data. Note WDB-2000-01, Centre for Biometry Wageningen, Wageningen.
- Dodd, KW (1996). A technical guide to C-SIDE. Technical Report 96-TR 32, Department of Statistics and Center for Agricultural and Rural Development, Iowa State University, Ames, Iowa. Available at http://www.card.iastate.edu/publications/DBS/PDFFiles/96tr32.pdf
- Dodd, K.W. Guenther, P.M., Freedman, L.S., Subar, A.F., Kipnis, V., Midthune, D., Tooze, J.A. and Krebs-Smith, S.M. (2006). Statistical methods for estimating usual intake of nutrients and foods: a review of the theory. *Journal of the American Dietetic Association*, 106: 1640-1650.
- Efron, B (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, **7**: 1-26.
- Efron, B & Tibshirani, RJ (1993). An introduction to the bootstrap. *Chapman & Hall*, New York.
- EFSA (2006). Guidance of the Scientific Committee on a Request from EFSA related to Uncertainties in Dietary Exposure Assessments. *The EFSA Journal* 438: 1-54.
- EFSA. (2009) Draft interim scientific report (unpublished).
- Ehrenberg, A.S.C., Uncles, M.D. & Goodhardt, G.J. (2004). Understanding brand performance measures: using Dirichlet benchmarks. *Journal of Business Research*, 57: 1307-1325.
- FAO/WHO (1997). Food consumption and exposure assessment of chemicals. Report of an FAO/WHO Consultation, Geneva, Switzerland. 10-14 February 1997.
- FAO (2002). Submission and evaluation of pesticide residues data for the estimation of maximum residue levels in food and feed. FAO, Rome.
- GenStat (2005). GenStat for Windows. Release 8.1, eigth edition, VSN International Ltd., Oxford.
- Goodhardt, G.J., Ehrenberg, A.S.C. & Chatfield, C. (1984). The Dirichlet: a comprehensive model of buying behaviour. *Journal of the Royal Statistical Society A*, 147: 621-655.

- Hamey, PY (2000). A practical application of probabilistic modeling in assessment of dietary exposure of fruit consumers to pesticide residues. *Food Additives and Contaminants*, 17: 601-610.
- Harris, C *et al.* (2000). Summary report of the International Conference on pesticide residues variability and acute dietary risk assessment. *Food Additives and Contaminants*, 17: 481-485.
- Harter, HL. Expected values of normal order statistics. Biometrika 48: 151-165.
- IUPAC (1995). Nomenclature in evaluation of analytical methods including detection and quantification capabilities (IUPAC Recommendations 1995). *Pure and Applied Chemistry* 67: 1699-1723.
- JMPR (1999, 2000). Reports of the joint FAO/WHO meetings of experts on Pesticide residues in food.
- Kipnis V, Midthune D, Buckman DW, Dodd KW, Guenther PM, Krebs-Smith SM, Subar AF, Tooze JA, Carroll RJ, Freedman LS (2009). Modeling data with excess zeros and measurement error: Application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics*, 65: 1003-1010.
- Kistemaker C, Bouman M and Hulshof KFAM (1998). De consumptie van afzonderlijke producten door Nederlandse bevolkingsgroepen - Voedselconsumptiepeiling 1997-1998. Zeist, TNO-Voeding (Report No: 98.812).
- Nusser SM, Carriquiry AL, Dodd KW & Fuller WA (1996). A semi-parametric transformation approach to estimating usual daily intake distributions. *Journal of the American Statistical Association*, 91: 1440-1449.
- Nusser SM, Fuller WA, and Guenther PM (1997). Estimating usual dietary intake distributions: adjusting for measurement error and nonnormality in 24-hour food intake data. In: Lyberg L, Biemer P, Collins M, DeLeeuw E, Dippo C, Schwartz N, and Trewin D (editors), *Survey Measurement and Process Quality*, Wiley, New York. p. 689-709.
- Odin, Y., Odin, N. & Valette-Florence P. (2001). Conceptual and operational aspects of brand loyalty. An empirical investigation. *Journal of Business research*, 53: 75-84.
- Pearson, ES and Hartley, HO. Biometrika tables for statisticians (1977). Vol II.
- Shimizu, K, and Crow, EL (eds). (1988). Lognormal distributions: theory and applications. Marcel Dekker, INC. New York.
- Slob, W., de Boer, W.J. and van der Voet, H. (2010). Can current dietary exposure models handle aggregated intake from different foods? A simulation study for the case of two foods. *Food and Chemical Toxicology*, 48: 178-186.
- Snedecor, GW & Cochran, WG (1980). Statistical Methods (7th edition). Iowa State University Click, Ames, Iowa.
- Souverein, O.W., de Boer, W.J., Geelen, A., van der Voet, H., de Vries, J., Feinberg, M. and van ’t Veer, P. (2011). Uncertainty in intake due to portion size estimation in 24-hour recalls varies between food groups. *Journal of Nutrition*, 141: 1396-1401.
- Tooze JA, Midthune D, Dodd KW, Freedman LS, Krebs-Smith SM, Subar AF, Guenther PM, Carroll RJ, Kipnis V (2006). A new statistical method for estimating the usual intake of episodically consumed foods with application to their distribution. *J Am Diet Ass.* 106: 1575-1587.
- Uncles, M.D., K.A. Hammond, Ehrenberg, A.S.C. & Davis, R.E. (1994). A replication study of two brand-loyalty measures. *European Journal of Operational Research*, 76: 375-384.
- van der Voet, H, de Boer, WJ & Keizer, LCP (1999). Statistical instruments for dietary risk assessment concerning acute exposure to residues and contaminants. Report August 1999, Centre for Biometry Wageningen, Wageningen.
- van der Voet, H, de Boer, WJ & Boon, P (2001). Modeling exposure to pesticides. Note HVT-2001-03, Centre for Biometry Wageningen, Wageningen.
- van der Voet, H. and Slob, W. (2007). Integration of probabilistic exposure assessment and probabilistic hazard characterization. *Risk Analysis*, 27: 351-371.
- van Dooren, MMH , Boeijen, I, van Klaveren, JD and van Donkersgoed G (1995). Conversie van consumeerbare voedingsmiddelen naar primaire agrarische produkten. RIKILT-report. Wageningen, RIKILT-DLO (Report No: 95.17).

- van Klaveren, JD (1999). Quality programme for agricultural products. Results residue monitoring in the Netherlands. RIKILT Institute of Food Safety, Wageningen.