BIKE manual

https://www.mdpi.com/2304-8158/10/11/2520

November 4, 2021

Contents

1 Introduction

Important note before using BIKE: always make a critical judgement of your data beforehand. Too little data or some features of data (e.g. censored values or lack of variation) can prevent proper estimation of model parameters. Also check how well the simulations of the model converge when running it.

BIKE is a model for foodborne exposure assessment, based on Bayesian methods. This manual should be read as an extension to the publication:

Ranta J, Mikkelä A, Suomi J, Tuominen P. BIKE: Dietary Exposure Model for Foodborne Microbiological and Chemical Hazards. Foods 2021, 10(11), 2520; https://doi.org/10.3390/foods10112520

For running the model, you need to install both OpenBUGS `http://www.openbugs.net/w/FrontPage` and R `https://www.r-project.org/`. It is recommendable to install R-studio. Within R, you need to install the following R-packages:

- R2OpenBUGS
- shiny
- mvtnorm
- shinyMatrix

Before using your own data, it is recommended to try BIKE with the synthetic example data provided. The example data sets are randomly generated in Excel and then saved as txt-files. One file is for concentration data, the other is for consumption data in tabular format. Additionally, two txt-files for other specifications are needed as explained later below. The tabular data need to consist of full tables (full column entries for each row). There can be additional columns as long as the necessary columns are provided with specific column names. Note that currently BIKE requires concentration data in a complete cross tabulation of all hazards and all food types. In practice, some hazard-food combinations may not occur in a given assessment task or they would pose practically zero exposures. However, in principle, any food could contain theoretically tiny amounts of any of the hazards (at least chemical), so that the possibility of true zero exposures might be excluded. For handling combinations that (assumingly) would not apply (or would not have data), one can enter a small set of dummy data and later select the adjustment factor as zero which then nullifies any exposure contribution from such source. Alternatively, one could fill in assumed concentration values drawn from expert opinions as "expert data" representing a hypothetical set of measurements.

Different data for test runs can be randomly generated from the same Excel file provided and then saved as a text file. The data sample size is made bigger or smaller by adding or deleting rows in the table. It is important to note that some minimum data always needs to be provided and a randomly generated play data does not necessarily always satisfy this. For example, all concentration data for each hazard need to contain at least some positive values above quantification limits.

2 Concentration data

Tabular data of hazard concentrations has to be in a file named `dataConcentration.txt` and it has to have at least these column names `Type, Hazard, Concentration, LOQ, LOD, Unit`. Each row contains one measurement result of one hazard from one food type. The `Type` is for the food name in question, e.g. 'broiler'. This must be the same food name for which consumption data are given in the other data file, so make sure it is spelled exactly the same way. There can be other columns giving broader food categories, e.g. 'poultry' or 'meat foods', but these are not used since the connection between hazard data and food consumption data is done at the finest feasible level of food type classification, as provided by user. Hence, `Type` can denote raw ingredients or composite food types containing many ingredients. The food names could be any character strings (without spaces), e.g. FoodEx2 codes or your own naming system, but very long names should be avoided for clarity and for more compact labels in the plotting windows. For example, 'minced meat casserole' could be shortened to 'mmeatcas' when preparing the data files.

Column `Hazard` specifies likewise the hazard name in question, e.g. 'cadmium' or 'salmonella' for each row. Column name `Concentration` is for the numeric concentration values measured for the specific hazard name and food type. Columns `LOQ` (limit of quantification) and `LOD` (limit of detection) specify the measurement limits. The notation format is the same for chemical hazards and microbiological hazards. The possibilities for each measurement are: reported numerical value ($>$ LOQ), a value between LOQ and LOD, or a value below LOD. The limits can also be different for each measurement. If a concentration value is reported in column `Concentration`, it is interpreted as an exact measurement. If the value was between the two limits, then both LOQ and LOD need to be given as numerical values, while concentration is marked NA. If the value was below LOD, then LOD needs to be given as numerical value, and both concentration and LOQ are marked NA. In this way, BIKE will know which of the three situations is in question for each hazard concentration measurement per row.

Note that although microbiological concentrations are often reported as "<LOD" or "<LOQ" the underlying microbiological interpretation is different. Due to the intrinsic counts of microbes they may appear zero in small analytical samples even when the concentration in larger volume (in the food unit or package or serving) is positive. Hence, the microbiological limits and concentrations are themselves estimates, not purely raw data. However, they are commonly reported and used as such. Original raw data should report the actual observations from dilution tube series or colony counts on petri dishes. Such data format is currently not implemented in BIKE.

The column for `Unit` is for specifying the measurement units, e.g. mg/kg, or cfu/g. These are not automatically converted to be compatible in calculations. Therefore, the user must use compatible measurements throughout when preparing the data files. If the concentration values are per gram, so must be the food consumption amounts as grams per day. A suitable measurement unit is such that it does not lead to extremely small or large numerical values since this could affect also numerical computations. BIKE will trust the user has selected sensible measurement units! Therefore, the column `Unit` is left for your own reference.

The concentration data entry alone does not specify whether the values below LOD should be interpreted strictly as small positives or possibly partially containing true zeros. In the latter case, a zero-inflated model would be used for estimating parameters of concentration distribution and pathogen

Samplenum	Food	**Type**	**Hazard**	**Concentration**	**LOQ**	**LOD**	Unit
1	poultry	broiler	cadmium	0.022	0.005	0.001	microgram.p.gram
2	poultry	broiler	cadmium	NA	0.005	0.001	microgram.p.gram
3	poultry	broiler	cadmium	NA	NA	0.001	microgram.p.gram
4	poultry	broiler	campylobacter	1.5	0.5	0.1	cfu.p.gram
5	poultry	broiler	campylobacter	NA	NA	0.1	cfu.p.gram
6	poultry	broiler	campylobacter	NA	0.5	0.1	cfu.p.gram
7	seafood	fish	cadmium	0.006	0.005	0.001	microgram.p.gram
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 1: Example of concentration data table. Required columns are shown bolded.

prevalence jointly from the concentration data. In the former case, the concentration data would represent only positive concentrations and hence a separate information is needed to estimate pathogen prevalence. The choice between the two interpretations (leading to two models) is specified in another input file with additional prevalence information given as surveillance samples if needed (see below).

3 Consumption data

Food consumption data corresponds to food diary data format where daily food consumption amounts per each individual are tabulated per food item, row-by-row. Tabular data has to be in a file named `dataConsumption.txt` and it has to have at least the column name `Weight` for the bodyweight and the column names for the detailed food types consumed on a specific day. For example, `broiler1` for consumption amounts of broiler on the first day. The next columns could be likewise `fish1`, `apple1`, etc. These columns would be followed by the same list of food types for the second day, e.g. `broiler2`, `fish2`, `apple2`, etc. There needs to be at least two days recorded for each consumer and the same number of days for all. Each row represents the reported consumptions of one consumer. The food types can represent composite foods or raw ingredients as needed, but the names of the food types (apart from the day number as the last character) have to be the same as those used in the hazard concentration data. Each row gives either the consumed food amounts, or zeros, for the reported days. The measurement units also need to be compatible with those in the concentration data, e.g. consumptions in grams if concentrations are given per grams. Consumption data may originally come in a hierarchical form that has several levels of food types with increasing details, e.g. seafood, fish, smoked fish, smoked salmon. However, the user should select only one of those labels (character string without spaces) which is used throughout in consumption data as well as in concentration data. This labeling of food items can only be as detailed as both data sets permit.

Idnum	**Weight**	**broiler1**	**fish1**	**broiler2**	**fish2**
1	64.0	0	169	107	0
2	78.6	38	80	0	205
3	88.9	0	100	95	0
⋮	⋮	⋮	⋮	⋮	⋮

Table 2: Example of consumption data table. Required columns are shown bolded.

4 Other required specifications

In addition to the above two data tables, other important information is needed. BIKE does not know which of the hazard names represent chemical or microbiological hazards. The file `occurrence.txt` needs to contain a table with the column names `hazardnames`, `hazardtypes` and rows for each hazard, specifying the name of the hazard (e.g. "cadmium") and the type ("K" for chemical, "M" for microbiological). The remaining columns have headers according to the food types, and the row entry will specify whether the concentration data for that food-hazard pair should be interpreted to represent only truly positive concentrations (even when below LOD), or as any measurements which might contain also truly zeros when the measurement fell below LOD. In the former case, the correct entry is "positives", and in the latter case the correct entry in occurrence table is "all". Note that this interpretation applies to full set of concentration data for the particular food-hazard pair.

The configuration also requires an entry for prevalence information per each food-hazard pair. The file `prevalencedata.txt` contains a table with column names `hazardnames`, `hazardtypes`, `infoods`,

`npositive`, `nsample`, and the row entries for the last two columns will define the number of true positives and the sample size to be used if the concentration data for that food-hazard pair should represent only positive concentrations. Otherwise, the number of true positives and the sample size are marked "NA". If the sample information is marked "NA", then the corresponding entry in occurrence table should be "all" to allow estimation of prevalence jointly with concentration distribution using zero-inflated modeling. Hence, make sure that "positives" in occurrence table goes together with specific values for prevalence sample data, and "all" goes together with "NA" in prevalence table.

Without the possibility to assign prevalence information in the form of surveillance sampling outcome, the concentration data table might become extremely large if it had to contain all zero entries (possibly thousands) among the possibly very few positive concentration measurements that were above LOQ. Also, it is fairly common that prevalence data and concentration data (for truly positives) are given separately, particularly with microbiological data which are often partially qualitative (presence / absence) and partially quantitative measurements. If all concentration measurements really represent truly positive values (even if below LOD) and the prevalence of positives should be assumed practically 100%, this can be implemented by setting `npositive` equal to a large value of `nsample` which forces the posterior estimate of prevalence towards 100%. However, this still leaves some small probability for the prevalence to be slightly lower than 100%. Without separate sample data the prevalence will be estimated from the concentration data using zero-inflated model which allows part of the values below LOD to be either true zeros or small positives.

These specifications in inputs determine how the data should be interpreted, which is an important assumption concerning the whole estimation of concentration distributions and prevalence. According to these specifications BIKE constructs the model to be used. Hence, the model always follows from those specified interpretations.

5 Driving BIKE

Once the input data files are prepared, BIKE can get started. The user interface is a shiny app, that can be started from R-studio by opening the file `BIKE.R` and running "run app" by clicking the tab. Note that the data files need to be in the same working directory as the R session, so change it if needed. BIKE will start by reading in the occurrence and consumption data and the additional specifications, after which it writes the BUGS-model code into a text file, to be run with OpenBUGS in the background. Some default results then appear in the shiny app window. After this, the user can select from a number of possible outputs that are processed from the now already simulated Markov chain Monte Carlo (MCMC) sample that resulted from the Bayesian model. Choosing other options can lead to a new simulation of the whole model. For example, there are two optional models for consumption frequencies. Choosing another model will construct a new BUGS code accordingly, and start a new simulation leading to new results. Also, choosing the number of MCMC iterations will start a new simulation. Large number of iterations can make computations slow, particularly in 2D simulations (needed for the quantile analysis of total exposure). It is recommended that several iterations of different lengths are tried to gauge how Monte Carlo error could still affect the results. There is no predetermined number of iterations that would guarantee sufficient accuracy because it depends on both model and data.

5.1 Selecting hazards and foods

From the selection tabs, the hazards and foods can be selected for which results are needed. However, the underlying model is running with the full set of hazards and foods, to be able to account for all pairwise correlations. The selection only specifies which results are processed as outputs. As a default the first hazard and the first food in the list is selected. A new selection does not re-start the whole MCMC simulation, but only processes the already existing MCMC output.

5.2 Concentration plots

Concentration plots for each hazard-food-combination can be selected from results. It is best to view them one-by-one for better visibility, although a few could be plotted simultaneously, according to the selected foods and hazards. These can be visualized either as probability densities, or as cumulative probability functions, either for absolute concentration values or log-values. A small Monte Carlo sample of variability distributions (magenta) is plotted to visualize the uncertainty about the variability distribution. The uncertainty distribution for its mean (gray) and median (black) is plotted in bold line. In log-scale, the mean and median are equal. For comparison with data, the raw data are represented as cumulative empirical distribution. If the data contain censored concentration values, two empirical distributions are plotted, one with lower bound substitution (blue) and one with upper bound substitution method (red). These represent the best case and worst case interpretations for censored values. Also, data points are plotted as tick marks on the x-axis, showing exact measurements in red, LOQ-values in green, and LOD-values in blue. Note that the concentration distributions in the figures represent truly positive concentrations.

Estimated proportion of true positives (i.e. hazard prevalence) is provided in a corner box. (Posterior median and 95%CI). Because the hazard prevalence is not a fixed assumption, it is always estimated with some uncertainty bounds. The additional data specifications will determine how the data are interpreted and which model is accordingly used. If prevalence sample data are given separately from concentration data, the prevalence is estimated based on the sample size and the number of positives in the sample. Hence, inserting a very large sample size with all positives will lead to prevalence estimate of nearly 100% with little uncertainty. In contrast, if concentration data represents all concentration, including possible true zeros, then it should be expressed in the additional specifications. This leads to a zero-inflated model which estimates prevalence simultaneously with other parameters for concentration distribution. Obviously, separate sample data are not used then. In both cases, the plot will always show distributions for positive concentrations and the estimated prevalence with 95%CI in a legend.

Note that the number of MCMC simulations affects the Monte Carlo accuracy of the results. Depending on data, the estimated distributions may have long tails which can make density plots very difficult to visualize informatively. Cumulative distributions often work visually better as shown in the example figure here.

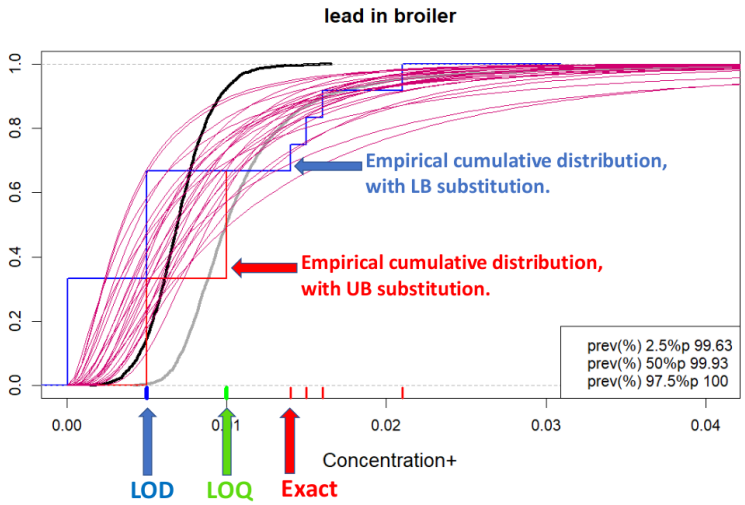

Figure 1: Cumulative distributions for separation of uncertainty and variability. The true variability distribution of positive concentrations is uncertain and this uncertainty is expressed by plotting a sample of probable variability distributions (magenta). The uncertainty distributions for mean (gray) and median (black) are shown too. Observed data for exact measurements are shown as red ticks ('Exact'-arrow points at the lowest quantified measurement). Also the limits LOD and LOQ are shown (blue, green). Here, approximately 35% of measurements were below LOD. Hence, the empirical cumulative distribution can start from zero (if LB substitution) or from LOD (if UB substitution) and this leads to two empirical cumulative distributions (blue and red) which are overlapping above LOQ. Estimated prevalence of contamination is given numerically with 95% CI and median. In this example, prevalence data was given (separately from concentration data) as a sample in which all sample units were positive. Hence, prevalence is estimated nearly 100%.

5.3 Consumptions plots

Consumption plots for single food types can be selected from the results, representing the variability distribution of servings (acute) and the variability distribution of mean servings (chronic), concerning actual consumption days. A 'serving' here means one day consumption. A mean serving (chronic) means a long term average serving for each consumer. The graphical distribution represents the mean servings as calculated from actual consumption days. These can be visualized either as probability density, or as cumulative probability, either for absolute concentration values or log-values. A small Monte Carlo sample of variability distributions (magenta) is plotted to visualize the uncertainty about the variability distribution. The uncertainty distribution for its mean (gray) and mode (black) is plotted in bold line. In log-scale, the mean and mode are equal. The numerical estimates for consumption frequency are provided in a legend. (Posterior median and 95%CI). Eventually, the consumption will result from the distribution of positive consumption amounts on consumption days, and the frequency of such consumption days. It is best to view distribution plots one-by-one for the selected foods for better visibility, although a few could be plotted simultaneously, as too many would not fit properly. For comparison with data, raw data are used to plot a cumulative empirical distribution (red). Also, data points are plotted as tick marks on the x-axis. Note that the number of MCMC simulations affects the Monte Carlo accuracy of the results.

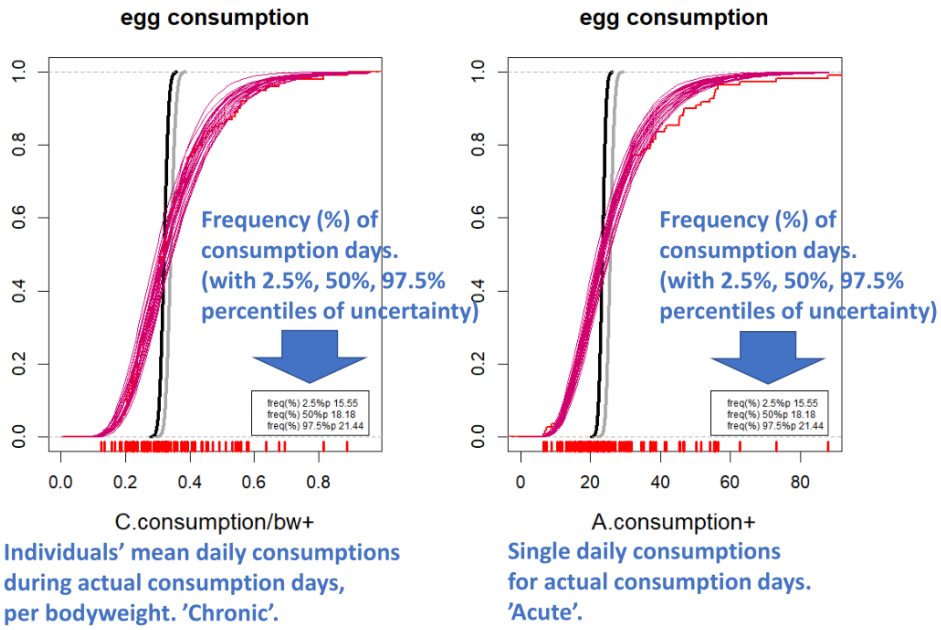

Figure 2: Cumulative distributions for separation of uncertainty and variability for both mean consumptions per bodyweight ('chronic') and single consumptions ('acute'). The true variability distribution of positive consumptions is uncertain and this uncertainty is expressed by plotting a sample of probable variability distributions (magenta). The uncertainty distributions for mean (gray) and median (black) are shown too. Observed data for positive consumptions are shown as red ticks and the empirical cumulative distribution in red. Estimated consumption frequency is given numerically with 95% CI and median.

5.3.1 Consumption model options

Consumption frequencies can be estimated using two alternative models. **'Independent days'** assumes the reported consumptions per day are from two or more mutually independent days. The consumption frequency is then estimated from a binomial model. **'Dependent days'** assumes the reported days are consecutive days so that the occurrence of consumption on any day may depend on the consumption on the previous day. The consumption frequency is then estimated from a Markov model as its stationary distribution. Changing the selection of model option leads to construction of a new model and new MCMC simulations.

If 'independent days' is selected for consumption frequencies, the model assumes two levels of variability: the variability of consumption days (yes/no) per consumer, and the variability of (logit) mean frequencies between consumers. However, some consumption data sets may contain food types that are reported as consumed on all days, for all consumers. Estimation of between consumer variability of the frequencies would not be possible then. Therefore, another option is offered for (un)selecting the between consumer variability. Both options can always be used, but obviously the estimation of between consumer variability does not produce meaningful results if such variability is not found in the data. Even if there were a few occurrences which manifest such variation, they may be too few for reliable estimation. For example, if all respondents (except one) in the data consume a food on all reporting days, and one consumes on 50% of days. And if the data contain only two reporting

days, the observable frequencies for any respondent could only be 100%, 50% of 0% per food type, which does not manifest a fine grained distribution of frequencies between consumers. Therefore, it is recommended to check the MCMC output for the variance parameter from the corresponding plots and to be aware of the quality of data used. If there are not enough data for estimating between consumer variance, it should be unselected from this model.

If 'dependent days' is selected for consumption frequencies, the Markov model assumes a common day-to-day transition probability for all consumers (no variability between consumers).

In all options, the distribution of consumption amounts is estimated for positive consumption days. The amounts and consumption occurrences are assumed independent so that e.g. consuming something rarely is not correlated with the amount when consumed.

5.3.2 Plots of correlated consumptions

The food consumption model contains two correlation structures for food types: all pairwise correlations for the actual consumption amounts in log-scale, and all pairwise correlations for the mean log-amounts. These are not specific correlations per each individual consumer, but generally for the whole group of consumers. These models aim to capture all such pairwise correlations in the consumer population if there are any. These can be inspected from pairwise plots of the food consumptions, for the selected foods. Although the plots show the selected foods only, the correlation models are always running for the full set of foods in the data. The plots show data points (blue stars) with model based simulated points (red) for each pair of food types. These can be drawn for the actual amounts or for the expected amounts. The plots are in log-scale. The artificial data for trying BIKE was generated with simplistic correlation structure which can be seen in the plots.

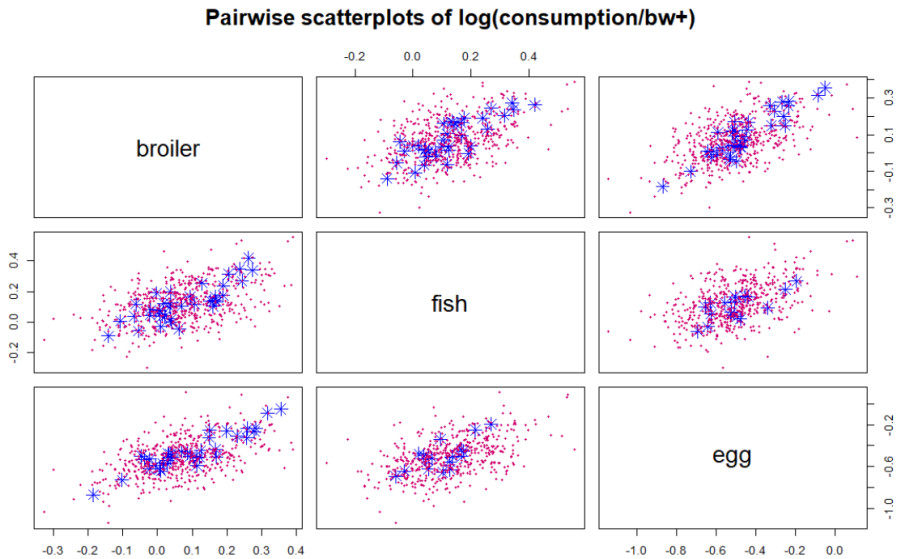

Figure 3: Pairwise scatter plots of positive food consumption amounts. Logarithms of consumptions per bodyweights. Data points (blue) and model based simulations (red).

5.4 Exposure plots

5.4.1 One hazard, one food type

Exposure plots for each hazard-food-combination can be selected from selection tabs. These are for assessing the exposure of one hazard from one food type. For the lack of space it is best to view them one-by-one for better visibility, although a few could be plotted simultaneously. These can be visualized either as probability densities, or as cumulative probability functions, either for absolute values or log-values of the exposure.

For microbiological hazards, the distribution of acute (individual single day) exposures is plotted, as concentrations per serving per day. Hence, a single 'serving' here means one day consumption. The plotted distribution is for actually contaminated foods on actual consumption days. For the microbiological exposures this provides the distribution of expected values of doses in the contaminated servings. The actual dose in a random serving would result as a random bacteria count from Poisson distribution that has such expected dose as the parameter.

For chemical hazards, the distribution of chronic (individual average) exposures is plotted, as average concentration per serving per day and bodyweight. The plotted distribution is for the individual mean exposures over actually contaminated foods on actual consumption days.

For both microbiological and chemical hazards, a small Monte Carlo sample of variability distributions (magenta) is plotted to visualize the uncertainty about the variability distribution. A variability distribution aims to describe how the positive exposure is distributed in the whole population of actual consumers. That is, the graphical image shows the worst case exposure that would result if all servings were contaminated and the food was consumed daily. The uncertainty distribution for the population mean (gray) and population median (black) of positive exposures is plotted in bold line. In log-scale, the mean and median are equal.

Visual comparisons between modelled exposure distributions and data are not exactly possible because there are no direct observations of exposures, only observed concentrations in food samples in one data, and observed consumptions of similar foods in another data. However, a pseudo empirical distribution of positive exposures can be simulated by randomly sampling (lots of) positive concentration values and positive consumption values from both data sets, and multiplying those values. The consumptions in such sampling are either the mean consumptions evaluated over actual reported consumption days or single consumptions from consumption days, depending on whether mean ('chronic') or single ('acute') exposures are needed.

Doing this pseudo empirical sampling once would provide one "empirical" exposure distribution from the data. This corresponds to a non-parametric exposure assessment model which samples directly from data values. To describe its uncertainty, the simulation is repeated with random versions of the same data sample, i.e. sampling with replacement. This corresponds to the bootstrapping method for uncertainty. Thus, a few "empirical" exposure distributions are plotted to visualize bootstrap uncertainty. If the concentration data contain censored concentration values, two versions of empirical exposure distributions are plotted, one with lower bound substitution (blue) and one with upper bound substitution method (red) for the concentration values. These represent the best case and worst case interpretations of censored values when simulating from data. The pseudo empirical LB and UB

simulations for exposure provide comparisons with the model based distributions (smooth curves), and all of these show the uncertainty in the form of a bundle of possible distributions.

All exposure distributions in the figures still represent truly positive exposures, i.e. when both the concentration of the hazard and the consumption of the food are positive. Since zero exposures are always possible (when either concentration or consumption or both could be zero for some days), the numerical estimate for the frequency of positive exposure days is provided in a legend. (Posterior median and 95%CI). This gives an estimate of how frequently the food is both contaminated and consumed. Thus, if the mean exposure due to actually positive exposures is estimated as 4 $mg/kg.bw$ and the exposure frequency is estimated as 50%, the resulting real mean exposure is 2 $mg/kg.bw$. A deeper analysis of the exposures from one or several foods is obtained from the quantile plots (see 'quantile estimation' below) where a specific population exposure quantile can be inspected. Note that the number of MCMC simulations affects the Monte Carlo accuracy of the results.

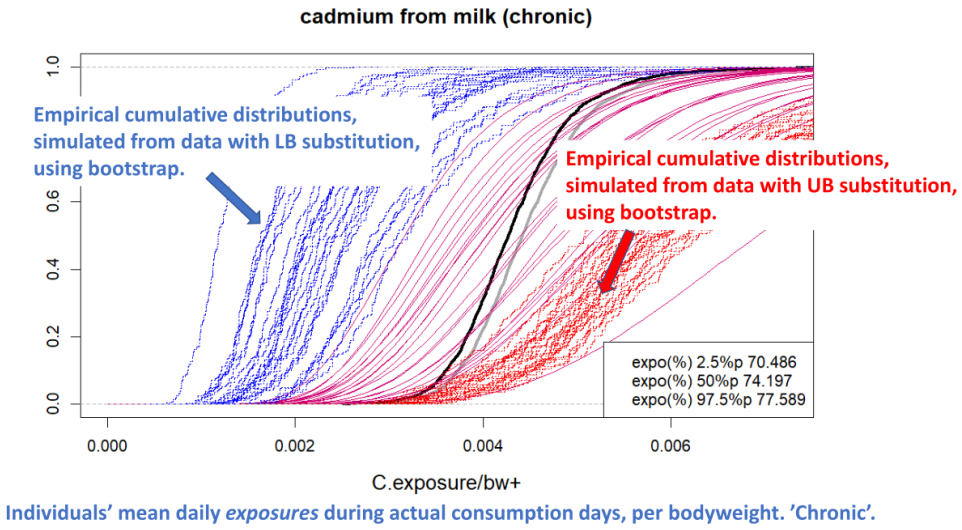

Figure 4: Cumulative distributions for separation of uncertainty and variability for mean exposures per bodyweight ('chronic'). The true variability distribution of positive exposures is uncertain and this uncertainty is expressed by plotting a sample of probable variability distributions (magenta). The uncertainty distributions for mean (gray) and median (black) are shown too. The empirical cumulative distribution simulated from bootstrapped data are shown using LB substitution (blue) and UB substitution (red) method. Estimated consumption frequency is given numerically with 95% CI and median.

5.4.2 Quantile estimation for summed exposure from several foods

The total exposure will be computed as the summed exposure from the selected foods. Depending on hazard type, the exposure will be either acute (microbiological) or chronic (chemical). From the distribution of total exposure, a selected quantile point is estimated. The quantile can be chosen from the selection tabs only when needed and the default choice is 'none'. The calculations employ 2D simulations and may be slow to compute. Therefore, they are produced only when requested. BIKE estimates the quantile for two exposure distributions: (A) the worst case exposure that results if all the foods are always contaminated and consumed daily, and (B) accounting the prevalence of contamination for each food as well as the actual consumption frequency. The worst case total exposure distribution is shown graphically (and numerically in a legend), while the actual total exposure quantile is given only numerically in a legend.

Positive exposure days only (worst case)

Positive acute exposure from a hazard-food combination occurs when both the hazard concentration and the food consumption are positive in a single serving day. **Positive chronic** exposure is likewise defined as the individual mean exposure from such positive exposures only. In other words: positive exposure does not include possible zero concentrations or zero consumption days. To evaluate total positive exposure from several foods, such positive exposures are summed over the selected foods. Note that this corresponds to a worst case scenario where all the selected foods are contaminated and they are all consumed on all days. The variation of such hypothetical total exposures is simulated to obtain a distribution for which the selected quantile point is estimated with 95% uncertainty bounds. To describe uncertainty about the variability distribution of such total positive exposures, a few distributions are plotted in the figure, and the uncertainty distribution for the quantile is overlaid between vertical bars showing the 95% uncertainty bounds (CI) for the quantile. The 95% CI and median are also given numerically in a legend.

The user can thus select the foods to be summed for exposure and the quantile to be investigated in the plot. Since the distribution of summed positive exposure can only be simulated approximately, the user can select the number of iterations to be used for producing a variability distribution, and the number of iterations to be used for producing the uncertainty distribution for the underlying n-tuple of parameters of each variability distribution. This is called 2D simulation of variability and uncertainty. More accurate results are obtained with more iterations, but this can become computationally slow. The 2D simulation for the quantiles does not require a re-run of the whole Bayesian model from the beginning, only a selected (thinned) subset of the existing MCMC sample of parameters and the chosen number of variability iterations per each parameter n-tuple in the subset.

Exposures for all days (including zeros)

Perhaps eventually of interest is the total exposure that accounts for both zero and positive daily exposures of the foods to be summed up. These occur according to the frequency of consumption days of each of the foods, and the prevalence of hazard contamination in each of the foods occurring on those consumption days. Typically, the proportion of zero exposure days can be large and some large exposure incidences may occur for some days only. This usually makes plotting of the total exposure distribution graphically infeasible. Therefore, the figure that appears in quantile plots shows a plot

representing the variability distribution of total positive exposures only (as explained above). This may serve as a worst case risk assessment tool. However, the quantiles for the overall total exposure distribution, including random zero exposures are nevertheless evaluated and numerically given in the other legend in the same figure. The estimated 95% uncertainty interval of the required quantile may become nearly entirely zeros if the positive exposures are rare occasions leading mostly to zero exposures. For proving a low risk, this would be a good result!

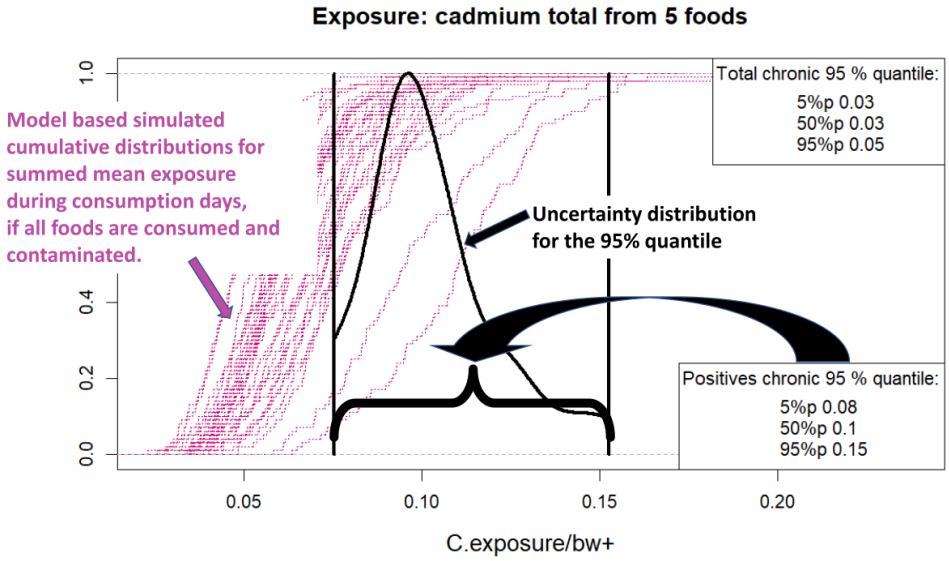

Figure 5: Cumulative distributions for separation of uncertainty and variability for mean exposures per bodyweight ('chronic'). The true variability distribution of positive exposures from all selected foods is uncertain and this uncertainty is expressed by plotting a sample of probable variability distributions (magenta). The uncertainty distribution for the selected quantile is shown between vertical bars. Estimated quantile is given numerically with 95% CI and median, for both the positive (lower corner) and all exposures (upper corner).

5.5 MCMC samples of parameters

The posterior distribution is computed using Markov Chain Monte Carlo (MCMC) sampling. This is an iterative sampling method which may need longer simulation runs than non-iterative Monte Carlo sampling. In MCMC, each iteration step depends on the values generated at the previous step. This sampling converges to the correct target distribution which is the posterior distribution of all unknown parameters. In difficulties, the posterior may be nearly flat or with multiple peaks. This, and/or insufficient iterations may then result into poor convergence, and the obtained parameter estimates would not be valid. Poor convergence and poor mixing of the algorithm could be detected as instability in the parameter distributions. The MCMC sample of the parameters, and the marginal posterior distribution are provided for critical inspection. Although a visual inspection is informal and indicative, it usually reveals problems in MCMC sampling. The sample should look smoothly scattered without clustering at some values only. The marginal probability density (plotted vertically on the left side) should look smooth without 'extra bumps'. New or increased MCMC runs should lead to same results. If a possible problem is detected, increase MCMC iterations and/or investigate whether there are simply insufficient data for the parameter to be well estimated. Note also that the full joint distribution is a multidimensional distribution which cannot be pictured and these plots only show marginal one-dimensional distributions for each parameter. The parameter samples can be viewed for both the concentration distributions and consumption distributions. The values at the peaks of those distributions could also be compared with the corresponding parameter estimates directly evaluated from raw data, e.g. mean and standard deviation, although censored data can make direct comparison less straightforward.

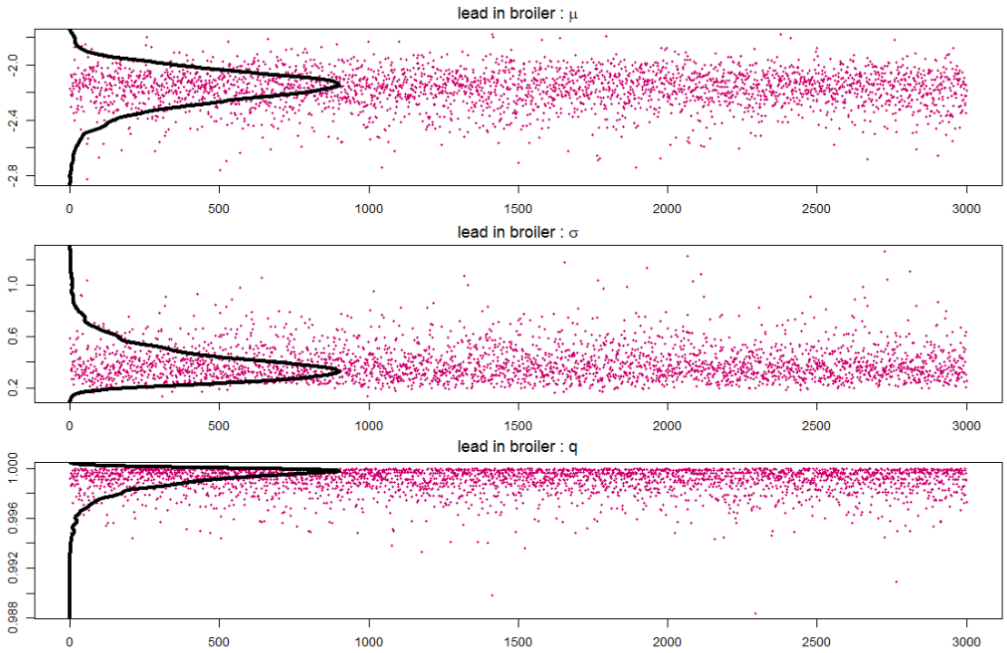

Figure 6: MCMC samples of model parameters μ and σ of the log-normal(μ, σ)-distribution, and prevalence q of the hazard. For each parameter, also the estimated marginal posterior distribution is shown sideways along the vertical axis.

5.6 Posterior predictive distributions tabulated

Posterior predictive distributions present predictions where all uncertainties and variabilities are integrated into one probability distribution. This can be a useful summary for assessing what is now probable, given all the data. The distribution can also be seen as the mean variability distribution, obtained by averaging over the uncertainty distribution of the parameters. This can be simulated for the concentrations, consumptions, and exposures.

5.7 Adjustment factors

Concentration data often do not describe contamination levels at the point of consumption but at some previous point in the food chain. This could be from retail level or from some stage of the production process, or even from raw materials. Some chemical contaminations could remain at those measured levels while others could be reduced by subsequent handling and cooking. Microbiological contamination levels are prone to large changes due to cooking (inactivation) or storage (growth/inactivation). A more detailed context dependent model would be needed to predict the final contamination levels. Such predictive models are not yet included in BIKE, but adjustment factors can be assigned for both the concentration level and prevalence. These will modify the modelled contamination when calculating exposure. The factor should be numerical but there is no automated check for the validity or sensibility of the factors given by user. A factor for prevalence should be between zero and one to ensure that the multiplication produces a mathematically valid prevalence. Likewise, a factor for concentrations should be a positive number. There is no uncertainty model of the factors, so they are simply given as constants. These could also be used for estimating the effect of hypothetical scenarios for intervention effects, if such would be assumed.

5.8 Some options

0 There are selection tabs for choosing between:

- Density plots and cumulative distribution function plots.
- Log-scale and absolute scale.
- Prior distributions for variance parameters: $\sigma^{-2} = \tau \sim$ Gamma(0.01,0.01) or restricted uniform distribution for σ.

The Gamma-prior is a conventional choice for inverse variance. The restricted uniform prior for standard deviation is determined from an exaggerated upper bound based on data.

There are also selections for the length of the whole MCMC simulation, and the number of variability iterations and uncertainty iterations for the 2D simulation of the quantile analysis.

5.9 Practical examples

5.9.1 Concentration of Cd in milk

Cadmium occurrence and milk consumption are generated in the synthetic data. The concentrations are assumed to represent truly positive concentrations (i.e. excluding true zeros). Hence, in the file DataOccurrence.txt the row for Cadmium at the column for milk has entry "positives". Correspondingly, in the file DataPrevalence.txt there is a row for the combination of Cadmium & milk with entries for sample size and sample positives to be used for prevalence estimation of true positive food samples. The example assumes 1000 positive samples out of 1000. This will lead to a prevalence estimate close to 100% with quite small uncertainty (result: 99.93% with 95%CI [99.63, 99.93]). You may replace this sample information as needed. However, numerical sample data entry here also requires that the corresponding entry in DataOccurrence.txt should be "positives". If numerical sample data are not available, it should be marked NA (both the sample size and the number of positives). In that case, the corresponding entry in DataOccurrence.txt should be "all" to denote that the concentrations values of the data table may be interpreted to contain all possible values, including possible true zeros when and if below LOD.

After BIKE has read the input data and started, the shiny interface should appear. From the selection tabs for Hazards to select one can click 'cadmium' and from the food types selection, 'milk'. The default choice for Scale is 'Absolute' and 'Cumulative' for distribution type. These can be changed in the left-hand panel.

Choose 'Concentration' under 'results to view' and from the lower part of the left-hand panel the desired number of iterations in MCMC simulations of model parameters. The default setting is only 4000 iterations which may not always give enough accurate approximations for results. When changing the setting for iterations, the output figures will be greyed until new results are computed.

From the cumulative distribution plot one can observe how uncertain is the estimation of true variability distribution of the Cd concentrations in the whole population of "milk foods". A bunch of smooth purple cumulative densities represent alternative distributions that are quite possible. The more dense is the bunch, the less uncertainty there remains where the true distribution lies. The true distribution of concentrations naturally has a mean and median, both of which are likewise uncertain. The uncertainty distribution for the mean is in gray, and for the median it is in black. Note that the graphical image shows the distributions for the truly positive concentrations only. The prevalence of such positive concentrations is estimated and shown in the corner box. The box shows the 95% uncertainty interval (i.e. Credible Interval, CI) with lower bound evaluated at 2.5% and upper bound evaluated at 97.5% of the uncertainty distribution. A useful point estimate for prevalence is evaluated at the 50% of the uncertainty distribution.

From the concentration plot, one can see that e.g. approximately 60% of the truly positive Cd concentrations in milk would be below 0.001. (The measurement unit is the one that was used in the input data). In the example this nearly coincides the proportion of measurements below LOD=0.001. Also, about 30% of the measurements were between LOD and LOQ=0.002. Only the remaining 10% were measured above LOQ, which is not much. Due to uncertainty, the 60% quantile is probably somewhere between 0.0005 and 0.0013. We cannot say that exactly 60% of concentrations are below 0.001. These judgements are only based on visual inspection of the bundle of cumulative distributions that are deemed most probable. A better 2D simulation of the quantile of your choice can be done using the quantile analysis panel.

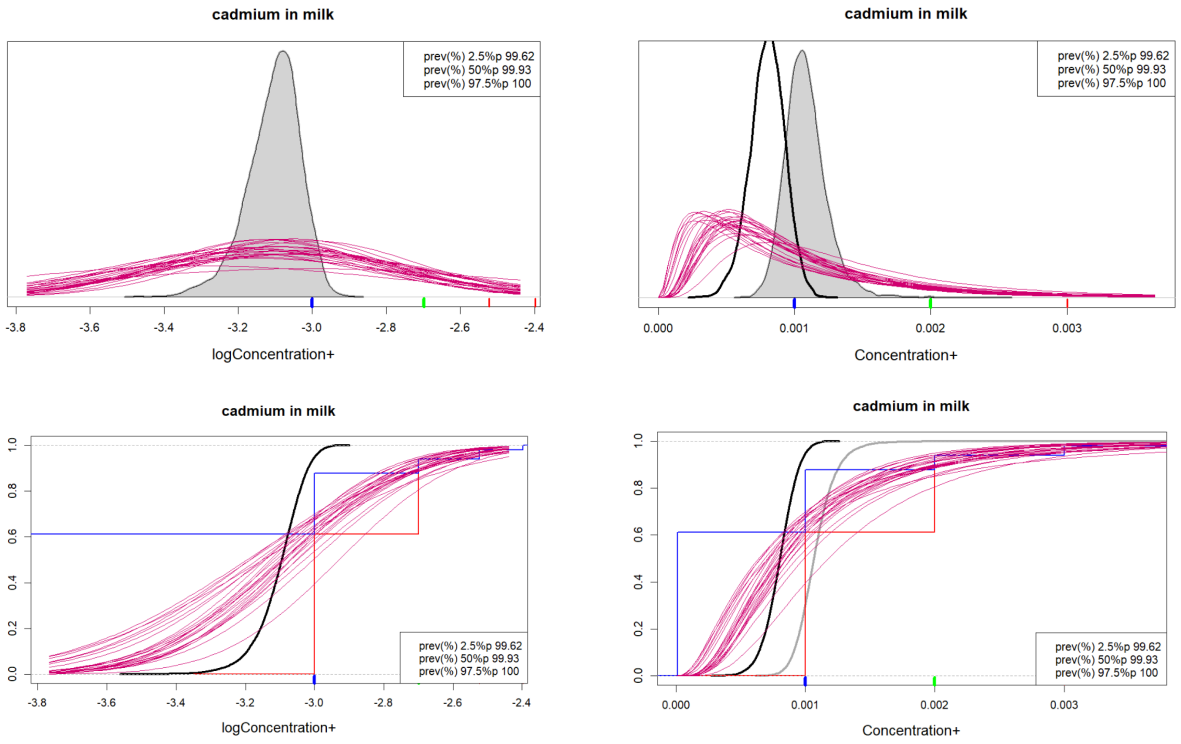

Figure 7: Plots for probability densities (upper) and cumulative probability functions (lower) for logarithmic (left) and absolute (right) concentrations.

5.9.2 Consumption of milk

The distribution plots for consumptions have similar options as the ones for concentrations. Except that the consumption measurements (as consumed amount on a reporting day) cannot be reported as below some "limits of detection", but either exact positive quantities or zeros. Hence, the distribution plot shows the distribution of non-zero consumptions. Consumption frequencies are estimated and shown numerically in a figure legend.

Depending on the dietary study, the reporting days could be completely independent days (as they often are) or they could be consecutive days which are not so independent: the consumption decision on the next day could depend on the consumption that took place on the previous day. Hence, there are two possible models for consumption frequencies: 'Independent days' and 'Dependent days' that can be selected. The latter could be useful for microbiological hazards in foods that are stored and consumed over several days. For Cadmium, a typical dietary data represent independent reporting days so this would be a proper option for estimating consumption frequencies. However, note that BIKE applies a multivariate normal model for the food consumption frequencies. This is in order to capture possible correlations between consumption frequencies of different foods. Even if results are only required for milk, the model is running for the full combination of foods in the background. Since chemical hazards are assessed for chronic (long term) exposures, the population distribution of individual mean consumptions per bodyweight is shown in the left frame of the figure. For your interest, the distribution of acute (random single day) consumptions is on the right. Both describe only those days when non-zero consumptions occur.

The model of milk consumption frequencies for independent days is based on estimating individual milk consumption frequencies and the between individual variations of those frequencies. This should account for the consumer variation that some individuals consume milk nearly on day-to-day basis, while others consume it less frequently. However, it is important to note that the estimation of such consumer differences requires that the data contain examples of some respondents who did consume milk more frequently in contrast to those who consumed less frequently during the dietary survey. If all respondents consumed milk on all reporting days, it would not be possible to estimate a variance parameter describing the variations in consumption frequencies between consumers. Including such variance parameter in the model would then be redundant and only leads to meaningless uncertainty about it. We need observed variations in order to estimate variation. Check if your data allows such estimation and see the resulting MCMC sample for the parameter in the diagnostic plots. If such variation cannot be feasibly estimated, one should select the corresponding option from the panel. Even when the reported consumption frequencies (as a pattern of zero and non-zero consumptions) differs between respondents, if there are only two reporting days per individual, the observable frequencies can only be 0%, 50% or 100%. The estimated frequency could still become e.g. 88% depending on relative proportions of those observations but some uncertainty prevails due to the nature of data. With more reporting days, the observable frequencies can obviously become more fine grained. Always know your data well to choose feasible estimation options!

The synthetic example data set was generated based on real consumption reports for the few food types from EFSA food consumption data base. However, since these were reported as univariate distributions per each food, it was not possible to recreate the true correlations between consumptions of different foods. This would naturally be contained in original food diary data. For the synthetic

data, some correlation was artificially added so that e.g. milk appears to be correlated with the other food consumptions. This can be inspected in scatter plots according to user selections. For the assessment of summed exposure from several foods, the correlations may be important because if each food would be simulated independently of the others, their total sum might add up unrealistically. Therefore, correlation structures can be relevant and these are included between the random actual amounts per day, as well as between the mean amounts of various foods.

5.9.3 Exposure of Cd from milk

The distribution plots for Cd exposure visualize the population distribution of chronic (long term mean) exposures in the population, among the actual exposures that were positive. That is: when the food servings were actually both contaminated and consumed. This can be graphically presented more easily than the distribution of exposures including the occasions when the food was either not contaminated or not consumed which would split the exposure distribution into a peak at zero and a long thin upper tail. Again, the exposure frequency is shown numerically in a legend. For Cd in milk, this example shows that contaminated milk is consumed on about 74.2 % of the days, 95% CI [70.6,77.6]. This is due to consumption frequency of milk because the prevalence of Cd in milk food was considered practically 100%. The lower the consumption frequency and the lower the hazard prevalence, the lower is the exposure frequency.

Likewise to previous images, the plot shows a bundle of chronical exposure distributions to visualize the uncertainty. For comparison, a bundle of cumulative distributions for the exposure are plotted using simulations directly from data values with lower bound (blue) or upper bound (red) substitution methods for concentration values below LOD or LOQ.

5.9.4 95% quantile of Cd exposure from one or several foods

It is often requested to estimate what is the exposure level L (from one or several foods) that 95% of the population has exposure less than L. This is the 95% quantile point. In chronic exposure assessment we look for long term mean exposure per bodyweight). Theoretically, if we knew the mean exposures for all individuals in the population, we could simply find out the 95% quantile point exactly from the sorted list of exposures. But since we have limited data we can only estimate the quantile point with some uncertainty. BIKE provides a selection of quantile points that may be estimated. It uses "2D Monte Carlo" simulation to simulate the uncertainty for the unknown parameters and the variability for the variables. Every variability distribution depends on some unknown parameters. When we increase the variability simulations, we get more correct variability distribution per each parameter choice. When we increase the uncertainty simulations, we get more correct uncertainty distribution for what those unknown parameters might be. We need both to get an accurate 95% uncertainty interval for the true 95% quantile point. This can become computationally slow, and the user is suggested to start with a modest number of iterations. The quantile results are provided for the total exposure from the selected foods; at minimum one food must be selected. Again, the graphical figure shows only the result for actually contaminated foods from actual consumption days. This is also given numerically in bottom-right corner. Eventually, the quantile point of interest should account also the prevalence of contamination and the frequency of consumption for each food item. This result is given numerically in top-right corner. Since many food items may typically be uncontaminated and are not consumed daily, this actual quantile point can become much lower than the worst case estimate

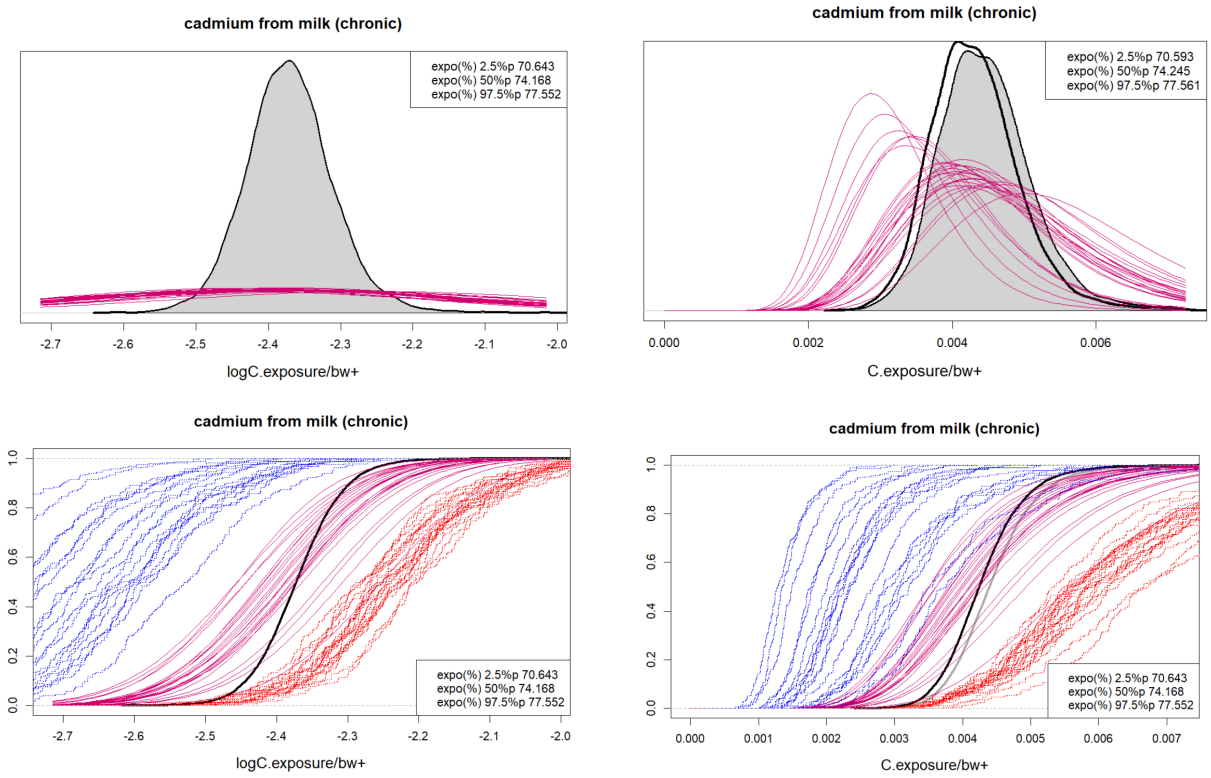

Figure 8: Plots for probability densities (upper) and cumulative probability functions (lower) for logarithmic (left) and absolute (right) positive chronic exposures per bodyweight. Exposure frequencies estimated in top-right and bottom-right legends. Cumulative probabilities are also given as simple Monte Carlo assessment with either lower-bound (blue) or upper-bound (red) substitution method for censored concentration values.

which only counts positive contaminations and positive consumption days. The example figure shows that the 95% exposure quantile resulting from five foods is [0.08,0.15] with 95% uncertainty, if all the five foods are always contaminated and consumed daily ("Positives chronic"). The eventual total 95% exposure quantile is only [0.03,0.05] when accounting for prevalences in contaminations and frequencies of consumptions ("Total chronic").

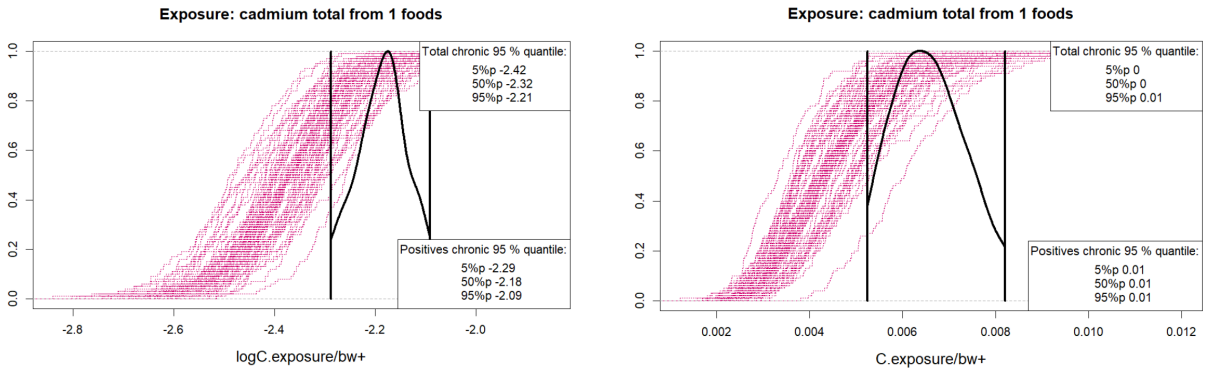

Figure 9: Plots of cumulative probability functions for logarithmic (left) and absolute (right) positive chronic exposures per bodyweight. Exposure from positive contaminations on actual consumption days estimated in bottom-right legend. Exposure accounting for contamination prevalence and consumption frequencies estimated in top-right legend. Uncertainty distribution (between vertical lines) for the 95% quantile point of exposure. This example shows the exposure from one food (milk) only.

5.9.5 Concentration of Campylobacter in broiler meat

In this example we need to bear in mind that the data are based on campylobacter concentrations in raw broiler meat samples. Therefore, to account for the effect of cooking or other processing and storage, one should apply the factors for prevalence and/or concentration based on some external research or expert opinion. Otherwise, BIKE only estimates the exposure as it would be according to the plain values drawn from data. For microbiological concentrations, it is most useful to plot log-values rather than absolute values. Therefore, log-scale option was selected. In the cumulative probability function plots it can be seen that nearly 50% of the concentration values were below LOD and nearly 80% below LOQ. From the empirical distribution alone it would be impossible to assess the distribution for low concentration values, whereas the model based continuous distributions show what the most probable distributions could be, based on both exact measurements and those between or below LOQ and LOD.

5.9.6 Exposure of Campylobacter from broiler meat

Referring to the previous example on concentration distributions, we note that the exposure distributions here also stem from the plain concentration values given in the data which only represent raw meat without processing factors. For microbiological exposure assessment, it is relevant to consider the acute exposure resulting from a single random serving (here a day). The distributions then could be used to assess whether the exposure is large enough to cause a risk of infection. Naturally, this depends on a dose-response assessment that is beyond the scope of exposure assessment in BIKE. From the results we could see the estimated frequency of positive exposure events (which require that a contaminated food unit happens to be consumed), and the distribution of absolute (or logarithmic) exposure in case such exposure event happens. The figures show cumulative probability functions for both logarithmic and absolute exposures per serving day. Again, pseudo empirical distributions are overlaid using lower bound (blue) and upper bound (red) substitution for concentration values. Note that the exposure is generated using Poisson distribution whose mean parameter is the product of concentration and consumption amount. This takes into account the randomness of bacteria counts

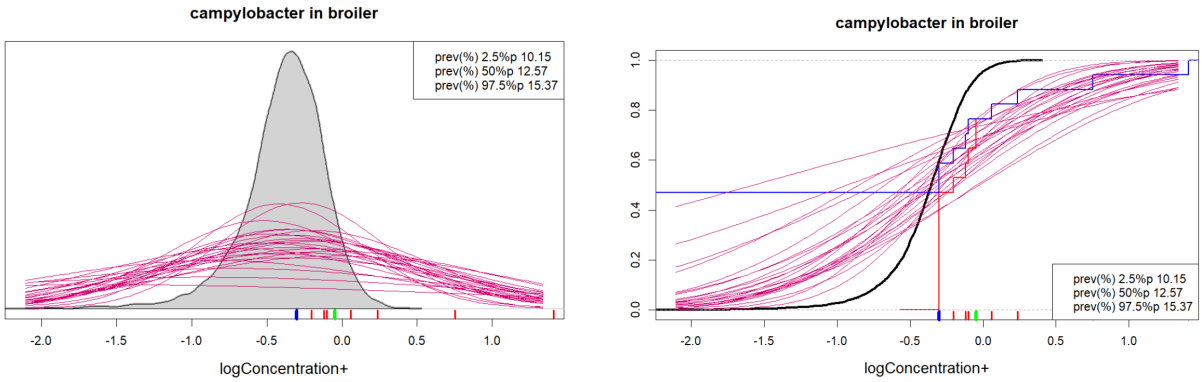

Figure 10: Plots for probability densities (left) and cumulative probability functions (right) for logarithmic Campylobacter concentrations (log-cfu/g). Estimated contamination prevalence is shown in legends. The cumulative probability plot also shows the empirical cumulative probability functions based on either lower-bound substitution (blue) or upper-bound substitution (red) for censored concentration values.

so that the actual dose is a discrete number of bacteria, $(0, 1, 2 \ldots)$.

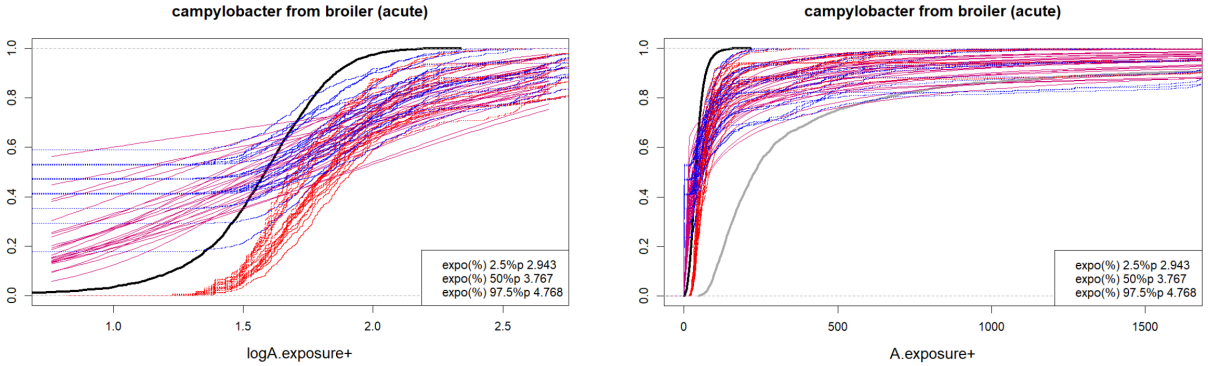

Figure 11: Plots for cumulative probability functions for logarithmic exposure (left) and absolute exposure (right) for Campylobacter from (raw) broiler meat (log-cfu or cfu per serving day). Estimated exposure frequency is shown in legends. The cumulative probability plots also show the empirical cumulative probability functions based on either lower-bound substitution (blue) or upper-bound substitution (red) for censored concentration values.

5.9.7 95% quantile of Campylobacter exposure from one or several foods

Referring to the previous example for cadmium, a similar assessment for Campylobacter exposure quantile from one food type (broiler) is presented in the figure. The logarithm of acute single dose has a variability distribution whose 95% quantile point is estimated to be [2.43,3.73] for contaminated servings on consumption days. When also the contamination prevalence and consumption frequency is accounted, the 95% quantile point is only $[-\infty,1.21]$ which shows that contamination and/or consumption does not always occur and a logarithm of zero exposure is always $-\infty$. The bundle of cumulative distributions show the uncertainty of the variability distribution for positive exposures. The left side of

the distributions have stepwise increments due to the discrete nature of bacteria counts. For example: two and three bacteria cells give $\log(2) = 0.301$ and $\log(3) = 0.477$.

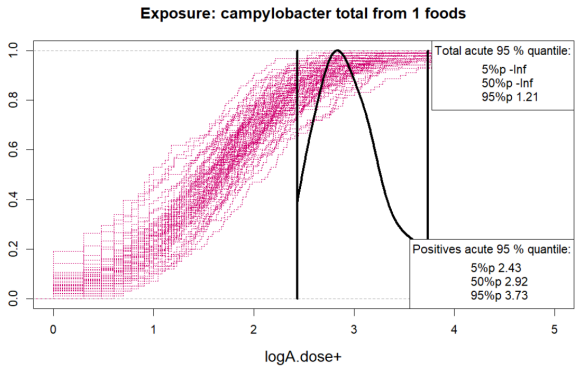

Figure 12: Plots of cumulative probability functions for logarithmic positive acute exposures per serving day. Exposure (Total acute) estimated in top-right legend. Exposure (Positives acute) from only contaminated foods on consumption days estimated in bottom-right legend. Uncertainty distribution (between vertical lines) for the 95% quantile point of positive exposures. This example shows the exposure from one food type (broiler) only.

6 Quick guide

BIKE model can be applied according to following steps:

- Download R, R studio and OpenBUGS software to your computer. Install also the R-packages R2OpenBUGS, shiny, shinyMatrix and mvtnorm.
- Save consumption data set (excel file in text format .txt) to the working directory of your choice by the name DataConsumption.txt. This file has to contain at least the following column headings:
  - Weight: Body weight of each respondent. Must be given for all individuals. Missing values not allowed.
  - Food type with reporting day: Name of the studied food items catenated with the index number of the reporting day. E.g Orange1 for the consumption of orange on day 1, and Orange2 for day 2, etc. There should be at least two reporting days for all respondents. These columns should contain consumptions of the food items for each respondent, e.g. grams per day. Note that the chosen weight units should match the units used for hazard concentrations. If food consumption is given as grams per day, the hazard concentrations in the food should be given as something per grams. BIKE does not convert measurement units. Take care they are throughout compatible. Also, the food names used in consumption data should match the food names used in concentration data.

  The number of reporting days must be the same for all persons. Missing values not allowed. Consumption data file may also contain other columns for your information although these are not currently used in the model. For example, age of respondents.
- Save concentration data set (excel file in text format .txt) to the working directory by the name DataConcentration.txt. This file has to contain at least the following column headings:
  - Type: Name of the food item, e.g. Orange. These should match with the names used in consumption data.
  - Hazard: Name of the hazard, e.g. Cadmium.
  - Concentration: Value of the measured concentration. If the measurement is below LOQ the value should be marked as NA.
  - LOQ: limit of quantification. If the measurement is below LOD the value should be marked as NA.
  - LOD: limit of detection.

  Concentration data file may also contain other columns for your information although these are not currently used in the model.
- Save occurrence file in text format DataOccurrence.txt to the working directory. The file contains the following columns that must be filled according to your data:
  - 1st column. 'hazardnames': list of the hazard names. E.g. Cadmium, Campylobacter.

  - 2nd column. 'hazardtypes': type of the given hazard, chemical "K" or microbiological "M".
  - remaining columns. Names of the food types for the food items represented in data. These columns specify how the concentration of the hazard (row) in this food type (column) should be interpreted in the case where concentrations are below LOD. Either they represent strictly "positives" or "all" values which could include true zeros. BIKE will choose a model according to this selection.

  Note that "all" for concentration information implies that the concentration distribution will be estimated jointly with prevalence parameter using a zero-inflated model where the fraction of measurements below LOD are interpreted allowing the possibility of true zeros. Then, prevalence data file should mark the corresponding hazard sample data as "NA" to signify the hazard prevalence is not estimated from separate sample information.

- Save prevalence file in text format DataPrevalence.txt to the working directory. The file contains the following columns that must be filled according to your data:
  - 1st column. 'hazardnames': list of the hazard names.
  - 2nd column. 'hazardtypes': type of the given hazard, chemical "K" or microbiological "M"
  - 3rd column. 'infoods': name of the food types in which the hazard occurs.
  - 4th column. 'npositive': number of detected positive samples. If sample information is not available, this should be marked NA.
  - 5th column. 'nsample': number of samples in total. If sample information is not available, this should be marked NA.

  Note that "NA" for sample information implies that the prevalence will be estimated jointly from the concentration data using a zero-inflated model where the fraction of measurements below LOD are interpreted allowing the possibility of true zeros. Then, occurrence data file should mark the corresponding hazard concentrations as "all" to signify they may contain both true zeros and small positive values when below LOD.

- Run the appBIKE.R file in R studio by clicking the 'run app'.

7 Notations

- **Concentration+**
  positive concentration (zeros excluded).
- **C.consumption/bw+**
  positive chronic (i.e. mean) consumption per bodyweight per consumption day (zeros excluded).
- **A.consumption+**
  positive acute consumption per consumption day (zeros excluded).
- **C.exposure/bw+**
  positive chronic (i.e. mean) exposure per bodyweight per exposure day (zeros excluded).
- **A.exposure+**
  positive acute exposure per exposure day (zeros excluded).
- **MCMC**
  Markov chain Monte Carlo sampling method.
- **uncertainty**
  uncertainty of parameter values (represented by posterior distribution, realized as an MCMC sample).
- **variability**
  variability of quantities in a population, modelled as a distribution that depends on its parameters.
- **quantile**
  quantile point of a variability distribution.
- **empirical distribution**
  a distribution of data values as such.
- **pseudo empirical distribution (of exposure)**
  a distribution of exposure produced by sampling concentrations and consumptions directly from the separate data sets for each. Usually requires either LB or UB substitutions for values below LOQ or LOD, leading to lower or upper estimate of pseudo empirical exposure distribution.
- **bootstrap**
  resampling of data (without replacement) to create artificial random replicate of data, with original sample size.
- **2D simulation**
  simulation of parameter values from uncertainty distribution (here by MCMC) and simulation of variable quantities from variability distributions (which are defined by parameters).
- **consumption frequency**
  proportion of actual consumption days in the long run.
- **prevalence**
  proportion of contaminated food items.