Risk Analysis, Vol. 31, No. 2, 2011 DOI: 10.1111/j.1539-6924.2010.01490.x

Nonparametric Predictive Inference for Exposure Assessment

V. J. Roelofs,* F. P. A. Coolen, and A. D. M. Hart

Exposure assessment for food and drink consumption requires the combining of information about people's consumption of products with concentration data sets to provide predictions for chemical intake by humans. In this article, we present a method called nonparametric predictive inference (NPI) for exposure assessment. NPI is a distribution-free method relying only on Hill's assumption $A_{(n)}$. Effectively, $A_{(n)}$ is a postdata exchangeability assumption, which is a natural starting point for nonparametric statistics. For further discussion we refer to works by Hill and Coolen. We illustrate how NPI can be implemented to produce predictions for an individual's exposure based on consumption, body weight, and concentration data. NPI has the advantage that we do not have to assume a distribution to implement it. There may, however, be information available to suggest a distribution for a random quantity. Therefore, we present an NPI-Bayes hybrid method where this information can be taken into account by using Bayesian methods while using NPI for the other random quantities in the model.

KEY WORDS: Bayesian methods; exposure assessment; *M* functions; nonparametric predictive inference

1. INTRODUCTION

There has been increased recognition among decisionmakers of the need to take account of uncertainty when considering risks to food safety and the environment. In an address to the European nonfood scientific committees, the European Community's Director General for Health and Consumer Protection said that: "Even though it is not a subject that lends itself easily to quantification, I would urge you to take account of the risk manager's need to understand the level of uncertainty in your advice and to work towards a systematic approach to this problem."[1]

Exposure assessments aim to evaluate the likely intake of substances by an organism. Many available methods rely on distributional assumptions.[2–4] In this article we suggest a new way of implementing exposure assessment using nonparametric predictive inference (NPI), which only relies on Hill's assumption $A_{(n)}$.[5] NPI provides lower and upper probabilities for the predicted value(s) of one or more future observation(s) of random quantities by including interval uncertainty. We illustrate the advantages of this method for this application, including the lack of distributional assumptions and the forming of bounds on censored data. We briefly present a new method where NPI is combined with Bayesian posterior predictive distributions for different random quantities in the exposure model. The NPI-Bayes hybrid method is useful in situations where we have enough information to assume a distribution for some random quantities in the model but do not want to assume a distribution for others. For example, there is frequently much less information available about concentrations of chemicals than about consumptions and body weights of a population. This

The Food and Environment Research Agency.
*Address correspondence to V. J. Roelofs, The Food and Environment Research Agency, Sand Hutton, York, YO41 1LZ, UK; tel: +44 (0)1904 462164; fax: +44 (0)1904 462111; victoria.roelofs@fera.gsi.gov.uk.

hybrid method also enables us to use NPI to see the effect of distributional assumptions in the Bayesian method.

In Section 1.1 we explain exposure assessment and the exposure model that we will use in this article. Section 2 introduces NPI. We explain NPI for exposure assessment in Section 3, where we also discuss how NPI deals with values that fall below a positive limit of detection (LOD). We illustrate NPI for a specific case study of young children's exposure to benzene from soft drinks in Section 4. Section 5 introduces the new NPI-Bayes hybrid method that we propose for combining random quantities modeled using Bayesian methods with random quantities modeled using NPI. This method is illustrated using an example in Section 6. Section 7 discusses the methods presented in this article and their usefulness in exposure assessment.

1.1. Exposure Assessment

The four main steps of risk assessment include hazard identification, effects assessment, exposure assessment, and risk characterization.[6] Here we focus on exposure assessment, which is the evaluation of the likely intake of substances. It involves the prediction of concentrations or doses of substances to which the population of interest may be exposed. Exposure can be assessed by considering the possible exposure pathways and the rate of movement and degradation of a substance. A simple exposure model that we consider throughout the article is:

$$\text{Exposure} = \frac{\text{Concentration} \times \text{Consumption}}{\text{Body weight}}, \quad (1)$$

where exposure is typically measured in μg/kg bw/day, concentration in μg/kg, consumption in kg/day, and body weight in kg. This model is often generalized by including different types of food and drink and the concentrations of the specific chemical of interest in each. For human risk assessment complicated exposure models are available,[7] where analysts are trying to combine different exposure pathways. However, as our aim is to explore the NPI methodology, we restrict attention to the simple model of Equation (1), where we only consider the exposure from food or drink represented by the single random quantity "consumption." From here on we will refer to Model (1) as the Exposure Model.

2. NONPARAMETRIC PREDICTIVE INFERENCE

NPI is a method that provides lower and upper probabilities for the predicted value(s) of one or more future observation(s) of random quantities. NPI is based on Hill's assumption $A_{(n)}$, explained in Subsection 2.1, and uses interval probability to quantify uncertainty.[8] It is an alternative to robust Bayes-like imprecise probability methods[9] with the advantage that it does not require the user to select sets of prior distributions. NPI has been presented for many applications, including comparison of proportions,[10] adaptive age replacement strategies,[11] and survival analysis involving right-censored data.[12] Due to its use of $A_{(n)}$ in deriving the lower and upper probabilities, NPI fits into a frequentist framework of statistics, but can also be interpreted from a Bayesian perspective.[13,14] Other advantages of NPI include that it is consistent with interval probability theory,[15] in agreement with empirical probabilities, exactly calibrated in the sense of Lawless and Fredette,[16] and it allows the analyst to study the effect of distributional assumptions in other methods. NPI makes only few assumptions, one of which is that the data are exchangeable, so we can put the data in order of magnitude irrespective of the order in which the observations were made. Exchangeability can be explained as follows. If we have n exchangeable random quantities, they are all equally likely to be the smallest, second smallest, etc. So, for any one of the random quantities, the probability of its rank among all these random quantities is uniformly distributed over the values 1 to n (assuming there are no ties for simplicity).

To our knowledge, NPI has not been used before in the area of exposure assessment, but as it can provide predictive probability bounds for the exposure of an individual without making distributional assumptions, it seems useful to explore its implementation. Therefore, we present an NPI analysis for exposure assessment using the simple Exposure Model (Section 1.1) in Section 3, to explain how it can be implemented and to illustrate the advantages of using a nonparametric method.

2.1. Hill's $A_{(n)}$

NPI is based on the assumption $A_{(n)}$, proposed by Hill[5] for prediction when there is very vague prior knowledge about the form of the underlying distribution of a random quantity, or if one explicitly wishes

not to use any such information. Let $x_{(1)}, \ldots, x_{(n)}$ be the order statistics of data $x_1, \ldots, x_n$, and let X_i be the corresponding random quantities prior to obtaining the data, so that the data consist of the realized values $X_i = x_i,\ i = 1, \ldots, n$. Then the assumption $A_{(n)}$ is defined as follows:[14]

1. The observable random quantities $X_1, \ldots, X_n$ are exchangeable.
2. Ties have probability 0, so $x_i \neq x_j$ for all $i \neq j$, almost surely.
3. Given data $x_i, i = 1, \ldots, n$, the probability that the next observation, X_{n+1} falls in the open interval $I_j = (x_{(j-1)}, x_j)$ is $\frac{1}{n+1}$, for each $j = 1, \ldots, n+1$, where we define $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$.

Hill's $A_{(n)}$ is a postdata assumption that one can make after the data have become available. $A_{(n)}$ avoids making strong assumptions about the data as there is no need to choose a probability distribution for the data. The assumption would require consideration if there were patterns in the data suggesting that postdata exchangeability is not appropriate. For example, if one sees a clear trend in the data, such as can occur in time series, the use of the inferences based on $A_{(n)}$ would not be appropriate. Clearly, if one has strong information about the probability distribution for the random quantity of interest and one is happy to use this information, then $A_{(n)}$ is not suitable as it explicitly avoids taking such information into account. In that case other methods may be more suitable, as we discuss in Section 5, where we combine NPI and the Bayesian posterior predictive distribution because there are differing amounts of information available for each of the random quantities.

For nonnegative random quantities, we define $x_{(0)} = 0$ instead and similarly if other bounds for the values are known. Hill's $A_{(n)}$ can be adjusted to include ties[13] by assigning the probability $\frac{c-1}{n+1}$ to the tied data point, where c is the number of times the value is present in the data set. In the NPI framework, a repeated value can be regarded as a limiting situation where the interval between the repeated observations is infinitesimally small, but can still be considered as an interval to which we can assign the probability $\frac{1}{n+1}$.

2.2. Lower and Upper Probabilities

We want to find lower and upper bounds for the probability that a future observation will fall in a particular interval. Then we can sum these probabilities to find lower and upper cumulative distributions for the value of the next observation. Here we illustrate how the lower and upper probabilities are calculated for the next observation X_{n+1}. More details can be found in Augustin and Coolen.[15]

We can find lower and upper bounds for the probability of $X_{n+1} \in B$ given the intervals $I_1, \ldots, I_{n+1}$, and the assumption $A_{(n)}$, where B is an element of $\mathscr{B}$ and $\mathscr{B}$ is the Borel σ-field over $\mathbb{R}$.[16] The Borel σ-field is the set consisting of all sets of intervals on the real line. B can be thought of as the union of any subset of the intervals in $\mathscr{B}$. The lower bound is then $L(X_{n+1} \in B) = \frac{1}{n+1}|\{j : I_j \subseteq B\}|$ and the upper bound is $U(X_{n+1} \in B) = \frac{1}{n+1}|\{j : I_j \cap B \neq \emptyset\}|$, where $|\cdot|$ denotes the number of elements in the set. The lower probability is the maximum lower bound for the probability that a future observation is in B corresponding to the assumption $A_{(n)}$. The $A_{(n)}$ assumption is necessary to allow us to assign equal probability of $\frac{1}{n+1}$ to each interval I_j, where $j = 1, \ldots, n+1$. The lower probability is the number of intervals totally enclosed in B divided by $n+1$ (the total number of observations plus one). Similarly, the upper probability is the number of intervals intersecting in any way with B divided by $n+1$. The NPI lower and upper cumulative distribution functions (cdfs) for X_{n+1} at x can then be calculated by taking $B = (-\infty, x]$. For a single random quantity these are step functions that increase by $\frac{1}{n+1}$ at each observation.

2.3. *M* Function

One useful tool for representing the probability mass on intervals for NPI is an M function, which is a basic probability assignment in the sense of Shafer.[17] This can represent the partial specification of a probability distribution on intervals with no restrictions as to where the probability mass falls in the interval although the masses must sum to one. For example, instead of a discrete probability mass function over the real line with probabilities for several points, an M function gives a probability mass that corresponds to an interval rather than a point value. This structure can be represented using the notation of the M function for a random quantity. The probability mass assigned for a random quantity X to an interval (a, b) can be denoted by $M_X(a, b)$. The intervals to which positive M function values are assigned can overlap. We use M functions later to represent the probabilities assigned to intervals by using NPI.

3. NPI FOR EXPOSURE ASSESSMENT

In this section we illustrate how to form NPI lower and upper cdfs for the Exposure Model. As left-censoring is a common occurrence in concentration data, we begin by discussing how NPI can deal with this censoring and then how to incorporate it into the resulting NPI lower and upper cdfs for exposure. In Section 4 we present an example to illustrate the method.

3.1. NPI for Left-Censored Data

When the concentration of a chemical is measured, there is often a positive LOD below which the equipment cannot measure reliably. The concentrations of the chemical that fall below the LOD will be recorded as "less than the LOD" rather than as a specific value. The censoring will be between 0 and the LOD because concentration cannot be negative. The distribution of probability mass for a data set with left-censored values can be represented using M functions, as explained in the following example.

Example: NPI lower and upper cdfs for left-censored data. Assume that the LOD is equal to 1 and that we have an ordered data set $\{x_1, x_2, x_3, x_4\}$ for concentration, where x_1 is a censored value between 0 and 1 and $x_i > 1$ for $i = 2, 3, 4$. The uncertainty about the value of x_1 can be represented by overlapping intervals in the M function, which takes into account that the censored value may be 0 or any value between 0 and 1. As we do not know the value of x_1, we can only say that for the next observation, x_5, there is a probability mass of $\frac{1}{5}$ on the interval $[0, 1)$. As the censored value could be anywhere between 0 and 1, the second interval in the M function, also containing probability mass of $\frac{1}{5}$, has to be $[0, x_2)$ as these are the tightest bounds we can have, given x_1 could be equal to 0. So the partial description of probability mass for a new observation X_5, based on $A_{(4)}$, can be represented as:

$$M_{X_5}[0\,,\ \ 1) = \frac{1}{5} \qquad M_{X_5}(x_3,\ x_4) = \frac{1}{5}$$
$$M_{X_5}[0,\ \ x_2) = \frac{1}{5} \qquad M_{X_5}(x_4,\ \infty) = \frac{1}{5}$$
$$M_{X_5}(x_2,\ x_3) = \frac{1}{5}.$$

The NPI lower and upper cdfs calculated from this M function will describe the tightest possible bounds given the information we have available and $A_{(4)}$. Note that using the previous definition, explained in Section 2.2, usually the intervals $(I_1, \ldots, I_{n+1})$ would be $[0, x_1), (x_1, x_2), \ldots (x_n, x_{n+1})$. However, here, as we do not know the value for x_1, the interval I_1 is $[0, 1)$ and the interval I_2 is $[0, x_2)$ as seen in the M function above. Fig. 1 shows the NPI lower and upper cdfs for this example.

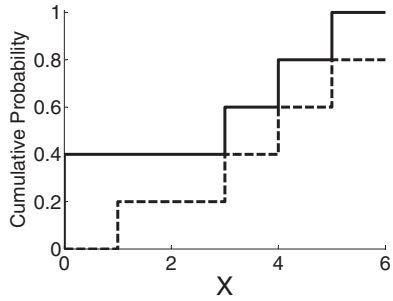

Fig. 1. NPI lower (dashed) and upper (solid) cdfs for X_5.

3.2. Example: NPI for the Exposure Model

Now we look at how to calculate NPI lower and upper cdfs for the Exposure Model that was described in Section 1.1. Assume we have observations for three independent random quantities X_i, Y_j, and Z_k (concentration, consumption, and body weight), $i = 1, \ldots, n_x$, $j = 1, \ldots, n_y$, and $k = 1, \ldots, n_z$. We have $n_x = 2$ ordered observations x_1, x_2, for the random quantities X_1 and X_2. Then we have intervals $(0, x_1), (x_1, x_2), (x_2, \infty)$. Assuming $A_{(2)}$, the probability that the next observation, X_3, falls in any of these intervals is $\frac{1}{n_x+1} = \frac{1}{3}$. Assume we also have $n_y = 3$ ordered observations y_1, y_2, y_3 for random quantities Y_1, Y_2, and Y_3, respectively. This leads to intervals $(0, y_1), (y_1, y_2), (y_2, y_3), (y_3, \infty)$ and assuming $A_{(3)}$, the probability that the next observation, Y_4, falls in any of these intervals is $\frac{1}{n_y+1} = \frac{1}{4}$. Taking the product of the intervals for the random quantities X_3 and Y_4 leads to 12 intervals, each with probability $\frac{1}{12}$, assuming there are no ties, for the random quantity that we call T_{new}. We combine the intervals by multiplying the minimum values of each interval for X_3 with the minimum values of each interval for Y_4 and the maximum values of each interval for X_3 with the maximum values of each interval for Y_4. Note that this works here because we are only using nonnegative random quantities. For example, to multiply (x_1, x_2) with (y_1, y_2), we multiply x_1 with y_1 and x_2 with y_2 to form the interval

$(x_1 y_1, x_2 y_2)$. Notice that this is the widest the interval can be, as combining the other endpoints, for example, y_2 with x_1, will always produce values that fall in this interval due to their ordering. Now assume we have $n_z = 2$ ordered observations z_1, z_2 for random quantities Z_1 and Z_2. Assuming $A_{(2)}$, the probability that the next observation, Z_3, falls in any of the intervals $(0, z_1), (z_1, z_2), (z_2, \infty)$ is $\frac{1}{3}$. Combining the intervals for the random quantities X_3, Y_4, and Z_3 as explained above in the Exposure Model leads to 36 intervals each with probability $\frac{1}{36}$ assuming there are no ties. The M function for the predicted value of the next observation, Exposure_{36}, is shown below, where we set $x_0 = y_0 = z_0 = 0$ and $x_3 = y_4 = z_3 = \infty$ and for all combinations of $i = 0, \ldots, 2$, $j = 0, \ldots, 3$, and $k = 0, \ldots, 2$. Note that $z_1 < z_2$ so $\frac{1}{z_2} < \frac{1}{z_1}$. The NPI lower and upper cdfs are formed as explained in Subsection 2.2.

$$M_{\text{Exposure}_{36}}\left(\frac{x_i y_j}{z_{k+1}}, \frac{x_{i+1} y_{j+1}}{z_k}\right) = \frac{1}{36}$$

4. CASE STUDY: BENZENE EXPOSURE

In this section we illustrate NPI lower and upper cdfs for the Exposure Model, using data for the exposure of young children to benzene in soft drinks. We have data for each of the three nonnegative random quantities, concentration, consumption, and body weight. A description of the data sets that we use for the analysis is given in Subsection 4.1. We calculate the NPI lower and upper cdfs for exposure for a random individual making the assumption that all three random quantities are independent. If it was believed that any of the random quantities were dependent, they could be combined to form one random quantity before applying NPI. For example, if body weight and consumption were thought to be dependent the ratio of consumption and body weight could be calculated and the random quantity "consumption ratio" could be used in the analysis instead. This loses some information as we no longer separate the random quantities, consumption, and body weight, but it does naturally include dependencies between these random quantities. To check the strength of the dependence between body weight and consumption, we considered Spearman's rank correlation coefficient, which ranges between -1 and 1. Values close to -1 indicate a strong negative rank correlation and values close to 1 indicate a strong positive rank correlation. Around zero indicates very weak correlation. Spearman's rank correlation coefficient is -0.007, indicating a very weak negative rank correlation between body weight and consumption. Therefore, we do not investigate using the consumption ratio here.

Table I. Concentration Data

Data Value	Frequency	Data Value	Frequency
<1	109	7	4
1	13	8	1
2	13	9	1
3	3	10	1
4	3	23	1
5	1		

4.1. The Data

Concentration data for benzene in soft drinks were obtained from the Food Standards Agency Survey from March 2006.[1] Out of 150 samples, 109 were below the LOD of 1 μg/kg. The concentration data are given in Table 1.

Assuming $A_{(150)}$, and denoting the ordered uncensored concentration values as conc_j, $j = 110, \ldots, 150$, the M function for concentration can be written as:

$$M_{\text{conc}_{151}}[0,\ 1) = \frac{109}{151}$$

$$M_{\text{conc}_{151}}[0, \text{conc}_{110}) = \frac{1}{151}$$

$$M_{\text{conc}_{151}}(\text{conc}_k, \text{conc}_{k+1}) = \frac{1}{151},$$

where $k = 110, \ldots, 150$ and $\text{conc}_{151} = \infty$. Notice that there are tied values in Table 1 and therefore if $\text{conc}_j = \text{conc}_{j+1}$ there will be probability mass at this point.

Consumption and body weight data were obtained from the U.K. Data Archive Study No. 3481, National Diet, Nutrition and Dental Survey of Children Aged 1.5–4.5 Years (1992–1993).[2] It is a four-day survey of 1,717 children giving information about their weight, food, and drink consumption and other covariates such as age, height, region, and social class. Only 1,694 individuals for whom complete data were available were used in the analysis. As an

[1] http://www.food.gov.uk/science/surveillance/fsisbranch2006/fsis0606.

[2] http://www.esds.ac.uk/findingdata/snDescription.asp?sn=3481&key=coding.

illustration of the use of NPI in exposure assessment we include both consumers and nonconsumers in the analysis and consider the average consumption over the four days of the survey. It should be noted that as with other dietary exposure models, the predictions are based on a small consumption survey data set, which only covers four days of the year. Predictions would be improved if larger databases of consumption were available to include in the analysis. We use the average consumption over the four days of the survey throughout this article, so for ease of presentation we will henceforth refer to this as "consumption."

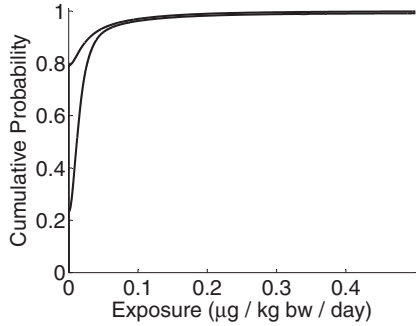

Fig. 2. NPI lower and upper cdfs for a random individual's exposure to benzene from soft drinks.

4.2. NPI Lower and Upper cdfs

In this example we derive NPI lower and upper cdfs for exposure for a random individual using the data sets that were described in the previous section. The lower cdf is the largest lower bound and the upper cdf is the smallest upper bound for the cdf corresponding to the $A_{(n)}$ assumptions. We assume all the random quantities are independent and calculate the NPI lower and upper cdfs for the exposure as described in Subsection 3.2.

To calculate the NPI lower and upper cdfs for exposure for a random individual, we first add a minimum and a maximum to each data set. For concentration and consumption, zero and ∞ are appropriate. However, for body weight we use 1 and ∞ to avoid the effect of dividing by 0. These lower and upper limits are chosen for illustration, any appropriate lower and upper limits could be used instead.

We have ties at 135 body weight values and 1,194 consumption values, both from an original sample size of 1,694. The tied values from the original data set and their frequency of occurrence are stored so the relevant probability can be assigned to the tied values themselves or to the intervals if the value is not tied. Working with the tied values speeds up computation and avoids problems with computer memory (computational issues are discussed in Section 4.3).

Fig. 2 shows the NPI lower and upper cdfs for exposure for a random individual. The NPI lower and upper cdfs appear to be smooth but are actually step functions that have small jumps at each calculated exposure value.

Notice that the results are initially a straight vertical line at 0, this is due to the nonconsumers included in the analysis. The large differences between the NPI lower and upper cdfs at low exposure values are due to the censored values in the concentration data. The final value of the lower cdf is given by:

$$\left(1 - \frac{(n_c + 1 + n_{cp})n_{bw}}{(n_c + 1)(n_{cp} + 1)(n_{bw} + 1)}\right) = 0.993, \quad (2)$$

as the probability mass on the interval between the highest exposure value and ∞ is $\frac{(n_c+1+n_{cp})n_{bw}}{(n_c+1)(n_{cp}+1)(n_{bw}+1)}$, where n_{bw} is the body weight sample size (here 1,694), and n_{cp} is the consumption sample size (here 1,694), and n_c is the concentration sample size (here 150). The NPI lower and upper cdfs provide bounds on the prediction of a random young child's exposure to benzene from soft drinks.

4.3. Computational Problems

In this subsection we briefly discuss some computational problems that arise when modeling NPI with large data sets. The large number of values in the consumption and body weight data sets described in Subsection 4.1 led to problems with computer memory as we needed to store all the possible combinations of all the values of all three data sets. As the data sets were $n_c = 150$, $n_{cp} = 1{,}694$, and $n_{bw} = 1{,}694$ in length (and adding a minimum or maximum depending on which cdf we consider) this leads to $(n_c + 1)(n_{cp} + 1)(n_{bw} + 1) = 433{,}826{,}775$ values for each cdf. These need to be stored along with the cumulative probabilities for each interval so that we can plot the NPI lower and upper cdfs. One way in which we solved this problem was by looking for repeated values in the data sets. Fortunately, when there are tied values it is possible to calculate NPI lower and upper cdfs for exposure for a random individual by only using the tied values once in the exposure calculation. The probabilities for each interval are calculated based on the number of repeated values that occur in the data sets. Eliminating repeated values

can be done at each stage so the calculation is only done with the minimum possible number of values.

If it is the case that even after checking for repeated values, the data sets are still too large, it is possible to calculate NPI lower and upper cdfs by counting how many values are less than various threshold values.

4.4. Imprecise Data

There are many practical situations where interval data may arise; some examples are given by Ferson *et al*.[18] These include cases where engineers and other scientists report the uncertainty associated with calibration of their measuring equipment with an interval and gross ignorance where we have no data about a random quantity so we assign theoretical limits. NPI can easily deal with this situation when there is an indication of the measurement uncertainty surrounding the values by assigning probability over wider intervals that take this measurement uncertainty into account. More details and an example are provided by Montgomery.[19]

5. THE NPI-BAYES HYBRID METHOD

Although NPI is a useful method and does not require distributional assumptions, there may be situations where analysts want to include additional knowledge about some random quantities in the model. In these situations an analyst may like to combine NPI results with Bayesian results. Therefore, in this section we explain the NPI-Bayes hybrid method for combining NPI and a Bayesian posterior predictive distribution. We focus on the important cases of the normal and lognormal distributions for the Bayesian method. However, it would be possible to apply this to other distributions by sampling from the corresponding posterior distribution using software such as Winbugs.[20] NPI focuses on predicting a future observation for a random quantity or a combination of random quantities. We therefore combine it with a Bayesian posterior predictive distribution, where a prediction is obtained for a random individual.

The Bayesian posterior predictive distribution for a future observation $\hat{y}$ is given by:

$$p(\hat{y} \mid \text{data}) = \int_{\theta} p(\hat{y} \mid \theta) p(\theta \mid \text{data})\, d\theta, \qquad (3)$$

where θ represents the parameters of the distribution and the data are assumed to be independent and identically distributed. For the normal distribution with a noninformative prior, $p(\mu, \sigma) = \frac{1}{\sigma}$, the Bayesian posterior predictive distribution is a Student t-distribution with location parameter $\bar{y}$, scale parameter $(1 + \frac{1}{n})^{\frac{1}{2}} s$, and $n - 1$ degrees of freedom.[21]

Assume that we have n_x observations x_i, where $i = 1, \ldots, n_x$, for random quantities X_i and that these observations come from a normal distribution. We also have n_y observations y_j, $j = 1, \ldots, n_y$, for positive random quantities, Y_j. As we have no further information about the Y_j we choose to use NPI for Y_{n_y+1}. To apply the NPI-Bayes hybrid method we assume independence between the X_i and Y_j.

For the X_i we invert the cdf of the Bayesian posterior predictive distribution for X at n_p equally spaced percentiles between 0 and 1 and assign each value probability $p_i = \frac{1}{n_p}$.

We want to find bounds on the prediction for the next observation, XY_{new}. To do this we use the following algorithm.

1. Take n_p values, which we denote $v_i, i = 1, \ldots, n_p$, by inverting the Student t-distribution with $n_x - 1$ degrees of freedom, location parameter $\bar{x}$, and scale parameter $(1 + \frac{1}{n_x})^{\frac{1}{2}} s_x$ at n_p percentiles.
2. Take the set of ordered observed values for Y_j, $j = 1, \ldots, n_y$, and form intervals (y_k, y_{k+1}), where $k = 0, \ldots, n_y$ and $y_0 = 0$ and $y_{n_y+1} = \infty$.
3. Multiply all the intervals by $v_i, i = 1, \ldots, n_p$ leading to intervals $(v_i y_k, v_i y_{k+1})$ for all i and with k as above. The probability on the intervals is $\frac{1}{n_p(n_y+1)}$.

We can describe the resulting probabilities using an M function where we set $y_0 = 0$ and $y_{n_y+1} = \infty$.

$$M_{XY_{\text{new}}}(v_i y_j, v_i y_{j+1}) = 1/n_p(n_y + 1)$$

for $i = 1, \ldots, n_p$ and $j = 0, \ldots, n_y$.

We illustrate the method for the (log)normal distribution; however, the Bayesian posterior predictive distribution for any distribution could be combined with NPI in a similar way if it is possible to sample from the Bayesian posterior predictive distribution.

6. USING THE NPI-BAYES HYBRID METHOD TO PREDICT EXPOSURE

We consider the simple Exposure Model where we let X, Y, and Z represent concentration,

consumption, and body weight, respectively. We begin by simulating a sample from a (log)normal distribution for each random quantity. Then we calculate exposure by combining NPI for some random quantities and the Bayesian posterior predictive distribution for other random quantities. We compare the results for each of these combinations. We will use the following notation for the different possible combinations: NX indicates that the NPI approach was used for the random quantities X_i and BX that the Bayesian posterior predictive distribution was used for the random quantities X_i. Similarly, we use NY, BY, NZ, and BZ.

6.1. Data Sets

To illustrate the NPI-Bayes hybrid method for calculating exposure, we need to have a sample for each random quantity in the model. In this example we choose distributions for each random quantity so that the data sets are similar to those from Section 4, which described young children's exposure to benzene in soft drinks. We simulate 20 concentration (x) values from a lognormal distribution with mean 1.4993 μg/kg and standard deviation 1.6749 μg/kg, 20 consumption (y) values from a lognormal distribution with mean 1.2776 kg/day and standard deviation 1.0159 kg/day, and 20 body weight (z) values from a normal distribution with mean 30 kg and standard deviation 3 kg. The ordered samples are:

X	0.1703	0.1828	0.3059	0.4278	0.4439	0.4994	0.5459,
	0.6037	0.6118	0.8656	0.8700	0.9074	1.175	1.471,
	1.472	1.569	2.346	2.663	4.036	12.12	
Y	0.2758	0.3199	0.4195	0.4397	0.4815	0.5377	0.6922,
	0.6997	0.7477	0.7675	0.8732	0.9954	1.130	1.174,
	1.348	1.403	1.629	1.632	1.653	2.769	
Z	25.86	25.93	26.58	26.65	26.93	28.61	28.83,
	28.98	29.37	30.95	31.11	31.51	32.12	32.18,
	33.11	33.57	34.66	35.59	35.87	36.34	

6.2. Calculating Exposure

We consider all different combinations of random quantities, using the NPI approach for some random quantities and the Bayesian posterior predictive distribution for the other random quantities in the Exposure Model. For ease of presentation, we display the results of each combination by the 10th, 50th, and 90th percentiles. These percentiles are either intervals, if they use the NPI approach for some random quantities, or point values if they only use the Bayesian posterior predictive distribution for all the random quantities. We represent these by plotting the intervals for each percentile in Fig. 3 as horizontal lines. To combine the distributions that the data were sampled from, we assume independence between the random quantities and call this the approximate exposure distribution. We show the percentiles of the approximate exposure distribution as vertical gray lines, so it is clear which intervals include the percentiles of the approximate exposure distribution. Although it cannot be seen in Fig. 3, the lower bound for the 10th percentile for the case (NX, NY, NZ) extends down to 0, and the upper bound for the 90th percentile for (NX, NY, NZ) extends to ∞.

As can be seen from Fig. 3, the 10th and 50th percentiles of the approximate exposure distribution do not lie within the corresponding lower and upper cdfs of any of the cases. However, the 90th percentile of the approximate exposure distribution lies within the lower and upper cdfs of all the cases except (BX, BY, BZ). The lower and upper cdfs on the 10th and 50th percentiles are all lower than the percentiles of the approximate exposure distribution. This is due to the samples used in this particular example. Using different samples would lead to different results where different percentiles are enclosed in the intervals as discussed in Montgomery.[19] The combinations where the NPI approach is used for more random quantities lead to the widest intervals. Therefore, if we were uncertain about the distributions that the data were sampled from, we would recommend combining all the random quantities using NPI because we are more likely to capture the percentiles of the approximate exposure distribution. The results from the NPI-Bayes hybrid method will depend on how representative the sample is of the true distribution and how well the chosen distribution for the random quantities modeled with Bayesian posterior distributions represent the distribution that the sample came from. As the sample size increases, the NPI lower and upper cdfs will always converge to the underlying distribution, whereas this is only true for the Bayesian posterior predictive distribution if the correct distribution type is chosen. It is also possible to include robustness to the prior distribution for the random quantities that Bayesian methods are used for.

7. DISCUSSION

In exposure assessment we want to quantify the uncertainty about exposure so we can produce robust

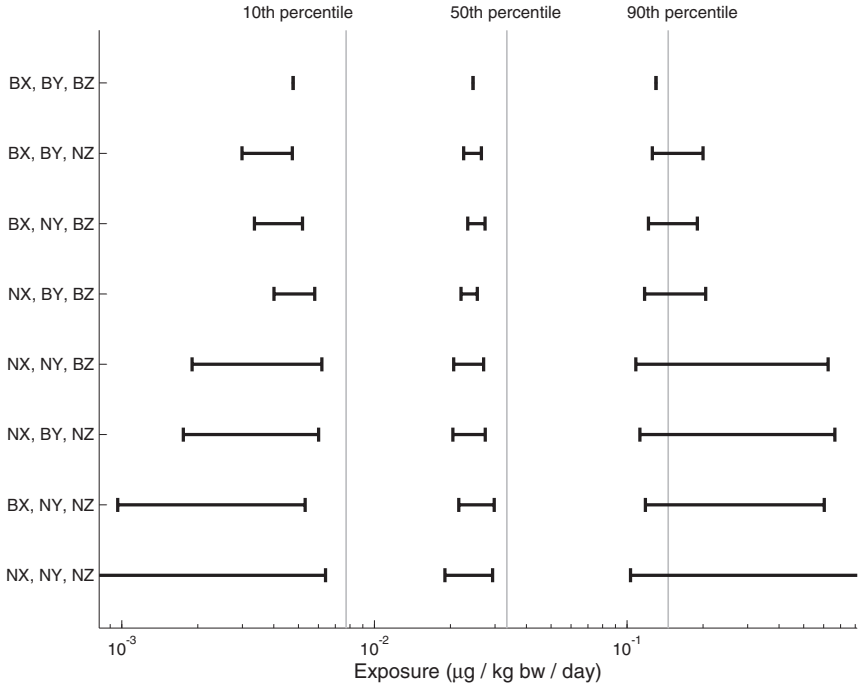

Fig. 3. Percentiles for exposure for combinations of Bayes and NPI for these specific samples of size 20. The vertical gray lines are the percentiles of the approximate exposure distribution.

predictions of exposure for individuals. As shown in this article we can use NPI for exposure assessment including censored data. The NPI lower and upper cdfs will always converge to the empirical distribution of the random quantity, which is not the case for distributional methods unless the distributional assumption is correct. As it is unlikely that the underlying distribution of the data is a standard distribution, for example, normal, gamma, beta distributions, distributional methods will not always be reliable whereas NPI only uses information from the data and therefore does not share this disadvantage. However, as NPI relies on data, it should only be used on medium-to-large data sets.

Exposure assessments often estimate a percentile of the population rather than producing a prediction for a single individual. In this article, we have only considered NPI for one future individual. NPI can be extended to $m \geq 2$ future individuals; however, care is needed as the random quantities corresponding to these m individuals are not independent. Coolen[4] explains how NPI can be used for multiple future observations and Arts *et al.*[22] present an application of such NPI for multiple future observations in the context of statistical process control. Considering m future individuals provides interesting opportunities to focus on specific proportions of these individuals for whom exposure levels may lead to risks, but increasing m will lead to increased imprecision, which restricts the practical value of this approach for values of m that are large compared to the number of available data.

We briefly discussed how to solve the computational challenges with implementing NPI either by only using one of each tied value in the sample or by using a threshold approach. We also illustrated how NPI can easily deal with censored data and fixed measurement uncertainty.

We explained how to calculate exposure when we have some random quantities where we assumed distributions in a Bayesian framework and others where we used NPI. This will be useful for situations where we only have prior knowledge about some random quantities in the model and we do not want to assume distributions for the other random quantities. It is common in practice that this situation, where we have lots of information about some random quantities and less information about other random quantities, will occur. For example, there is often little information about concentration of chemicals in different food types but lots of information available about the body weights of the population.

We have shown that NPI can be used to describe uncertainty about the percentiles of exposure for an individual without making strong distributional

assumptions and can be combined with Bayesian methods if required. It is a data-driven method and therefore we would recommend that it is used for medium-to-large sample sizes as are common in exposure assessment. As we have illustrated, the method is easy to use and could be implemented for more complicated models than illustrated here. For example, multiple food types could be taken into account or individuals could implement NPI to calculate their own exposure based on their own consumption and body weight. If random quantities were believed to be dependent on one another they could be treated as one random quantity in the NPI method.

ACKNOWLEDGMENTS

The first-named author has been funded by an EPSRC-CASE studentship supported by the Food and Environment Research Agency (Defra Seedcorn Funds). We would also like to thank Phil Northing from Fera for providing the consumption data and the reviewers for their helpful and thoughtful comments.

REFERENCES

1. Madelin R. The Importance of Scientific Advice in the Community Decision Making Process. Opening address to the Inaugural Joint Meeting of the Members of the Non-Food Scientific Committees. Directorate General for Health and Consumer Protection, European Commission, Brussels, 2004.
2. Chatterjee A, Horgan G, Theobald C. Exposure assessment for pesticide intake from multiple food products: A Bayesian latent-variable approach. Risk Analysis, 2008; 28(6):1727–1736.
3. Allcroft DJ, Glasbey CA, Paulo MJ. A latent Gaussian model for multivariate consumption data. Food Quality and Preference, 2007; 18(3):508–516.
4. Montgomery VJ, Coolen FPA, Hart ADM. Bayesian probability boxes in risk assessment. Journal of Statistical Theory and Practice, 2009; 3(1):69–83.
5. Hill BM. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. Journal of the American Statistical Association, 1968; 63(322):677–691.
6. Van Leeuwen CJ, Hermens JLM (eds). Risk Assessment of Chemicals: An Introduction. Dordrecht: Kluwer Academic Publishers, 1995.
7. Brand E, Otte PF, Lijzen JPA. CSOIL 2000 an Exposure Model for Human Risk Assessment of Soil Contamination. A Model Description. Bilthoven, The Netherlands: RIVM Report, 2007.
8. Coolen FPA. On nonparametric predictive inference and objective Bayesianism. Journal of Logic, Language and Information, 2006; 15(1–2):21–47.
9. Walley P. Statistical Reasoning with Imprecise Probabilities. London: Chapman and Hall, 1991.
10. Coolen FPA, Coolen-Schrijner P. Nonparametric predictive comparison of proportions. Journal of Statistical Planning and Inference, 2007; 137(1):23–33.
11. Coolen-Schrijner P, Coolen FPA, Shaw SC. Nonparametric adaptive opportunity-based age replacement strategies. Journal of the Operational Research Society, 2006; 57(1):63–81.
12. Coolen FPA, Yan KJ. Nonparametric predictive inference with right-censored data. Journal of Statistical Planning and Inference, 2004; 126(1):25–54.
13. Hill BM. De Finetti's theorem, induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference (with discussion). Pages 211–241 in Bernardo JM, DeGroot MH, Lindley DV, Smith AFM (eds). Bayesian Statistics 3. Oxford: Oxford University Press, 1988.
14. Hill BM. Parametric models for An: Splitting processes and mixtures. Journal of the Royal Statistical Society B, 1993; 55(2):423–433.
15. Augustin T, Coolen FPA. Nonparametric predictive inference and interval probability. Journal of Statistical Planning and Inference, 2004; 124(2):251–272.
16. Lawless JF, Fredette M. Frequentist prediction intervals and predictive distributions. Biometrika, 2005; 92(3):529–542.
17. Shafer G. A Mathematical Theory of Evidence. Princeton, NJ: Princeton University Press, 1976.
18. Ferson S, Kreinovich V, Hajagos J, Oberkampf W, Ginzburg L. Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty. Albuquerque, NM: Sandia National Laboratories Technical Report SAND2007-0939, 2007.
19. Montgomery V. New statistical methods in risk assessment by probability bounds. Ph.D. thesis, Durham University, 2009. Available at www.npi-statistics.com. Accessed on May 7, 2010.
20. Spiegelhalter DJ, Thomas A, Best N, Lunn D. Version 1.4.1. BUGS. Medical Research Council (MRC) UK, WinBUGS, 1996–2004. Available at: http://www.mrc-bsu.cam.ac.uk/bugs. Accessed on August 13, 2009.
21. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis. London: Chapman and Hall, 1995.
22. Arts GRJ, Coolen FPA, van der Laan P. Nonparametric predictive inference in statistical process control. Quality Technology and Quantitative Management, 2004; 1(2):201–216.