HOW TO HANDLE NON-DETECTS?

Ronald Hoogerbrugge and A.K. Djien Liem
Laboratory for Organic-analytical Chemistry, National Institute of Public Health and the Environment, P.O. Box 1, 3720 BA Bilthoven, The Netherlands

Introduction

In the determination of dioxins in environmental and biological samples the results are usually expressed in toxic equivalents (TEQ) of 2378-TCDD. The calculation of this TEQ value requires concentrations for all components with a toxicity equivalency factor (TEF). In practice however the concentrations of some components may be below the limit of detection (LOD). The appearance of non-detects in a particular data set is often a major problem for the assessment of statistical parameters (e.g. mean and standard deviation) or to assess the comparability of the data (e.g. concentrations of PCDD/Fs in human milk across Europe). The problem of non-detects is expected to grow especially in the field of PCDDs and PCDFs, because levels in foodstuffs and in environmental samples seem to decline[1].

Several imputation strategies have been evaluated on the basis of an artificially censored (by increasing the LOD) data set of dioxin concentrations in cow's milk (1989-1990). This "old" data set has only few non-detects and is therefore very suited for comparing the real measurement values with the values from the various estimation strategies. Also strategies to estimate the uncertainty in the calculated TEQ values will be tested.

Imputation approaches

In order to estimate the TEQ in the presence of non-detected data several strategies for imputation of the non-detects have been tested.

A) replace each non-detected element by zero

B) replace each non-detected element by the detection limit

C) replace each non-detected element by the half of the detection limit

In applying option C and assuming that: each non detected concentration is uniformly distributed between 0 and LOD and that: concentrations of several components are independent, the variance due to each non-detected component is $LOD^2/12$. The total variance in TEQ due to the non-detects is the sum of TEFs squared weighted variances of the non-detected concentrations.

D) When a sample in a data set has a non-detected value for a component the non-detect is replaced by the minimum of usual contribution to the TEQ and the LOD. The uncertainty due to the imputation process is estimated using the law for the propagation of error on the variation in the contributions.

E) Multiple imputation[2] with censoring of data. Like D the multiple imputation uses a model to impute values on the position of the non-detect. To reflect the uncertainty the imputation does not use the model value itself but chooses a random value from a normal distribution around the model value. The standard deviation of this distribution is equal to the residual standard deviation between the model and the measured values. The model for each congener is generated using Multiple Linear Regression. In the resulting iteration process the levels of the first congener are modeled using the concentrations of other congeners. Then the concentrations of the second congener are modeled and imputed and so on. After imputation of the last congener the iteration is continued with re-imputation the first congener. Then the process is repeated for the first which a Monte Carlo Markov Chain (MCMC). In the iteration process imputation candidates above the LOD are rejected and re-sampled. Mathematical details of the multiple imputation procedure will be presented elsewhere[3]. For this study 4 of such imputed data sets are generated.

Diagnostics and software

To evaluate the performance of the various imputation methods three general parameters are calculated for each approach and shown in table 1. The *systematic deviation* is the average difference between the estimated TEQ value and the TEQ value based on the complete data set. The *standard deviation of prediction (sd)* is the square root of the average of the squared difference between the estimated and complete TEQ values. This value can be interpreted as the standard deviation of uncertainty of the imputed values. The *relative standard deviation of prediction (rsd)* is the square root of the average of all squared differences divided by their estimate of uncertainty. This parameter gives an indication weather the estimated uncertainty gives an appropriate measure of the uncertainty for the particular data set. For an appropriate estimator of the uncertainty in the TEQ this value should be close to 1.
The calculations were performed using Matlab5 [4]. For this purpose some dedicated programs (m-files) are written. Calculation of imputation options A- D are relatively simple and are also possible in spreadsheets etc. Programming of the multiple imputation is much more dedicated. These calculations are also more time (5-20 minutes CPU) consuming.

Data set

The original data set consists of concentrations of the 17 toxic PCDDs and PCDFs in 303 individual samples of cow's milk[5,6]. The majority of the samples was taken from dairy farms in the vicinity of municipal waste incinerators. This data (from 1989-1990) were selected since recent sets of dioxin in cow's milk data contain more results below the detection limit and are therefor less useful as a reference set. From this original data set a censored copy was made by raising the LOD from 0.1 to 1.0 pg/g fat. Then the number of non-detects increases from 17 to 60%.

Table 1. Average systematic deviation (sys.dev.), standard deviation of prediction (SD) and standard deviation relative to its estimate (RSD) for the artificially censored data. The results are shown for the several imputation strategies (A to E) described in the text.

Code	Imputation method of non-detects	sys.dev.	SD	RSD
A)	zero	-0.68	0.75	2.53
B)	LOD	0.82	0.94	3.49
C)	0.5 LOD	0.07	0.24	1.00
D)	Contribution with censoring	0.16	0.21	1.27
E)	Multiple imputation	0.03	0.14	0.65

Results and Discussion

All results are summarized in Table 1. Imputation of zero (A) gives a negative systematic deviation and of the LOD (B) a positive one. Compared to the use of the lower and upper bound estimates the imputation of half the detection limit shows a remarkable improvement. The error of prediction (sd) reduces by a factor of 3-4. Also the estimated standard deviation of uncertainty seems to describe the differences between the estimated and reference TEQ values quite well (the rsd is not very different from 1). Figure 1shows the difference between the estimated and reference TEQ-values with the calculated standard deviation intervals. For samples with low levels (1-4 pg TEQ/g fat) mainly positive deviations dominate while for samples with higher levels (4-8 pg TEQ/g fat) the negative deviations occur. This might cause problems in trend or other comparison studies. For the highest levels (>8 pg TEQ/g fat) hardly any deviations are found because then the number of results below the LOD is small.

For method D the systematic deviation is very small compared to the other single imputation methods studied and also the error of prediction is reasonably well described by its predictor (rsd is close to 1).

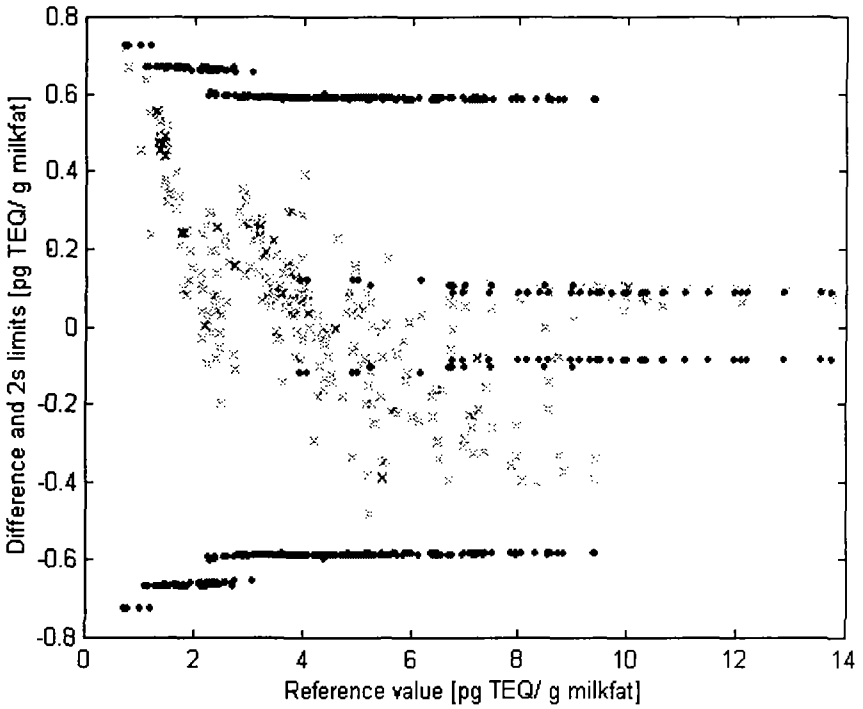

Figure 1) Difference between the TEQ estimates and the reference estimate plotted against the reference TEQ estimate for the data set with concentrations below 1 treated as 0.5 LOD (x) compared to the calculated 2 standard deviation limits (.).

Table 2 shows that the multiple imputation result has a small systematic deviation and a small scatter compared to the single imputation methods. The estimate of the rsd is smaller than 1 which indicates that the estimate of the standard deviation of uncertainty is not underestimated. Figure 2 shows that also no concentration dependent trends are present in the differences indication that these results can be very well suited for *trend* analysis, *epidemiological studies* etc.

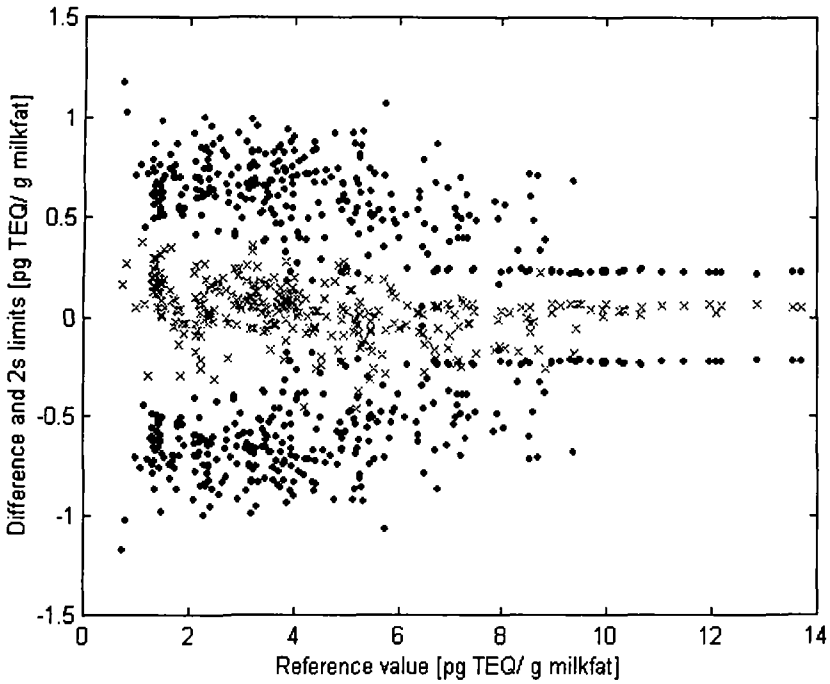

Figure 2) Difference between the average of the multiple imputation TEQ estimates of the artificially sensored data set and the original (reference) TEQ estimate of data. For comparison the calculated 2 s (standard deviation) limits of the multiple imputation data are shown (as dots) indicating that the uncertainty in the average is not underestimated.

Conclusion

Imputation of half the detection limit (C) yields an acceptable estimate of both the TEQ and its associated standard deviation of uncertainty. Imputation of the average contribution gives a slight improvement of this estimate (D). For this data set the most accurate estimate is generated by the multiple imputation algorithm (E). For each strategy the proposed uncertainty estimate adequately describes the difference between the imputed values in the censored set with the reference values.

References

1 European Commission Directorate General Health and Consumer Protection 2000. Scientific Co-operation of Questions relating to Food – Assessment of dietary intake of dioxins and related PCBs by the opulation of EU Member States. Report of SCOOP Task 3.2.5, May 2000
2 D.B. Rubin 1987. Multiple Imputation for Nonresponse in Surveys. New-York: John Wiley.
3 R. Hoogerbrugge and A.K.D. Liem in preparation
4 The Mathworks, Inc. http://www.mathworks.com
5 A.K.D. Liem and R.M.C Theelen Thesis "Dioxins: Chemical Analysis, Exposure and Risk Assessment" (1997) University of Utrecht.
6 A.K.D. Liem, R. Hoogerbrugge, P.R. Kootstra , E.G. van der Velde, A.P.J.M. de Jong, Chemosphere, Vol. 23, Nos.11-12, pp 1675-1684, 1991.