did not consider in depth the role of high percentages of imputed values in case of measurements below the LOD for the usefulness of Benford’s law in evaluating occupational hygiene data and that they did not explain the rationale behind their selection of data sets from the vast range of substances and countries that supplied data for the ExAsRub project.

REFERENCES


Benford’s Law and the screening of analytical data: the case of pollutant concentrations in ambient air. Analyst; 130: 1280–5.


doi:10.1093/annhyg/meu009
Advance Access publication 18 February 2014

REPLY

Author’s Reply to Koppisch et al. 2014

Frank de Vocht

Centre for Occupational and Environmental Health
The University of Manchester

Hans Kromhout

Institute for Risk Assessment Sciences
Utrecht University
E-mail: Frank.deVocht@manchester.ac.uk
Submitted 15 January 2014; revised version accepted 15 January 2014.

We would like to thank Dr Koppisch and co-authors for their interest in our paper in which we described the use of Benford’s Law (BL) for occupational hygiene data, using two data sets from the EXASRUB database (de Vocht et al., 2005) as illustrative examples (de Vocht and Kromhout, 2013). We could indeed have used other data sets collated within the ExAsRub data set but included the n-Nitrosamines data set from the MEGA database because of its interesting feature that although a standard number of specific n-Nitrosamines were measured only those below the limit of detection (>LOD) were reported in the MEGA database. To make the n-Nitrosamines data useful for exposure assessment within the EXASRUB project, we had to ‘manipulate the data’ by adding the missing (<LOD) concentrations for each measurement. Moreover, because of the
size of the data set \(N = 18,619\), we avoided problems resulting from small sample variability that would make its use as an illustrative example difficult.

Indeed, as pointed out by ourselves and confirmed by Koppisch et al. (2014), BL can be used to evaluate the quality of data on occupational exposures and if the data do not adhere to BL, there are a number of reasons why this could be, including a significant number of values imputed with a single value (e.g. values below the LOD in this particular data set). If we adhere to the four possible explanations outlined in de Vocht and Kromhout (2013), Koppisch and colleagues and we agree that “There is a reasonable explanation for the deviation from Benford’s law” (explanation number 3). In contrast to Koppisch et al., however, who state that in their opinion compliance with BL cannot be used as a criterion for data manipulations in data sets with a high percentage of values below the LOD, we would argue that evaluations using BL identified manipulation of the MEGA-ExAsRub \(n\)-Nitrosamines data set (e.g. imputation by LOD/\(\sqrt{2}\)) which, for example, provides the argument to use an alternative imputation method when the data are used in the future, such as, for example, described by Lubin et al. (2004). Nonetheless, we would like to thank Koppisch et al. for the additional theoretical work they included in their letter further supporting the results from our work by providing indications of the percentage of imputed values at which deviations of BL start to occur.

With respect to the queries about the numbers, the correct number of imputed specific \(n\)-Nitrosamines measurements in the MEGA database is the 13 376 eluded to in Koppisch et al. However, somewhere in the transfer and handling of the data between the conversion from MEGA to MEGA-ExAsRub part of these values were replaced by 0.10 (e.g. the LOD) but not registered as missing data (hence the resulting \(N = 6739\)). Indeed, we should have explained this in more detail in our manuscript. We double-checked the original ExAsRub work to ensure that indeed these values were also replaced by LOD/\(\sqrt{2}\) (e.g. \(\sim 0.07\)) prior to the analyses in de Vocht et al. (2007). However, from the data presented by Koppisch et al., it seems the number of specific \(n\)-Nitrosamines measurements from the MEGA database originally made available to the EXASRUB project \(N = 5243\) (de Vocht et al., 2007) and the number (currently?) available in the MEGA database \(N = 8564\) differ significantly, while some of the measurements in our pre-imputation data also includes values of 0.10, which are very likely values <LOD added as LOD (to avoid confusion, these are not the imputed values mentioned above, but the original provided data). This most likely explains the observed differences. Nonetheless, we thank Koppisch and colleagues for pointing out two errors in our Table 2. The \(R\) values were accidentally calculated using base(e) instead of base(10) and should have been 3.2, 4.5, 3.4, and 4.3 making them borderline acceptable for evaluation with BL (correspondingly, for clarity, equation 4 in our manuscript should have been log(10) as well). Furthermore, the \(P\) value associated with the 1 BL chi-square test for the BRMA rubber process dust data \((\chi^2 = 47.92 (df = 8))\) was indeed an error and should have been \(P < 0.001\), but this has no impact on our conclusions since the data sets remain to deviate significantly from 1 BL with the \(\chi^2\) values for the MEGA and MEGA-ExAsRub \(n\)-Nitrosamines data much larger than those for the National Exposure Database and British Rubber Manufacturing Association rubber process dust data.

We are encouraged to see that evaluation using BL has led to additional evaluation of, and discussion about, a data set of occupational exposure measurements—exactly how our paper was intended.

REFERENCES