# DATA HARMONIZATION

We provide detailed examples of data harmonization for data from an epidemiological cancer study (Stenehjem *et al*., 2015.1), a biomarker study (Testa *et al*., 2005.2), and an experimental animal study (Li *et al*., 2006.3).

Stenehjem *et al*.1 is part of our primary set of studies that directly investigated the exposure-response relation between benzene and AML incidence. The authors examined the incidence of lympho-haematopoietic cancers due to benzene exposure among 24,917 male offshore oil workers followed during 1999-2011. The authors provide hazard ratios (HRs) and 95% confidence intervals (95%CIs) for three categories of cumulative benzene exposure in ppm-years when compared to unexposed workers. These can be directly used to calculate estimates for the logRR and its standard error by log-transforming reported HRs and 95%CIs as is customarily done for meta-analyses. The (3x3) covariance matrix for the combined set of logRRs was estimated using the approach outlined by Greenland and Longnecker (1992)4 and implemented in the *covar.logrr* function from the *dosresmeta* R package. To use this function, the user is required to provide additional details on the number of cases and the total number of subjects or person-years of follow-up, which were available from the paper by Stenehjem *et al.*1. Data harmonization steps for studies included in the leukemia set were the same as those used for the AML set.

Testa *et al.*2 studied chromosomal aberration frequencies (a cancer biomarker) in 62 benzene exposed workers. Benzene exposure was monitored using stationary sampling and expressed as a time-weighted average (TWA) exposure level. The prevalence of chromosomal aberrations was assessed in (at least) 200 metaphases and reported as the proportion of aberrations. For the controls this was 1.08% (sd: 0.81%) and for exposed workers 2.52% (sd: 1.58%). Cumulative benzene exposure was estimated by multiplying the reported average exposure for exposed workers with the duration of exposure. A logRR was estimated by subtracting the log-transformed average proportion of aberrations in unexposed workers (controls) from that in exposed workers. The SE was estimated by applying the delta rule to standard errors of the reported average proportions to account for the log-transformation and standard variance rules for the difference between two independent random variables. Data harmonization steps for other studies in the CA and MN sets were similar. For the (few) studies that included multiple exposure levels, the full covariance matrix of logRRs was estimated using the same approach as for AML and leukemia.

Li *et al.*3 exposed 10 mice to 300 ppm benzene for 6 hours/day, 5 days/week, for 26 weeks. Six out of 8 animals in the unexposed and 8 out of 10 animals in the exposed group developed lymphoma’s in haematopoietic and lymphoid tissues. Cumulative exposure was estimated by multiplying the benzene exposure level by the exposure duration (in years). The RR [95%CI] was estimated to be 107 [0.64, 1.77] using standard relative risk calculations and were converted to a logRR and SE using the same approach as used for the AML and leukemia studies.

Table S1. Results from the data harmonization step for three selected studies.

|  |  |  |  |
| --- | --- | --- | --- |
| **Authors (year)** | **Cumulative benzene**  **exposure (ppm-years)** | **logRR** | **Covariance matrix** |
| Stenehjem *et al.* (2015) | 0.02 | 0.34 |  |
|  | 0.08 | -0.16 |
|  | 0.54 | 1.58 |
|  |  |  |  |
| Testa *et al.* (2005) | 43 | 0.85 |  |
|  |  |  |  |
| Li *et al.* (2006) | 150 | 0.06 |  |