# Estimation of the Exposure–Response Relation between Benzene and Acute Myeloid Leukemia by Combining Epidemiologic, Human Biomarker, and Animal Data

Bernice Scholten[1,2], Lützen Portengen[1], Anjoeka Pronk[2], Rob Stierum[2], George S. Downward[1], Jelle Vlaanderen[1], and Roel Vermeulen[1]

Downloaded from http://aacrjournals.org/cebp/article-pdf/31/4/751/3116574/751.pdf by guest on 26 September 2023

## ABSTRACT

**Background:** Chemical risk assessment can benefit from integrating data across multiple evidence bases, especially in exposure–response curve (ERC) modeling when data across the exposure range are sparse.

**Methods:** We estimated the ERC for benzene and acute myeloid leukemia (AML), by fitting linear and spline-based Bayesian meta-regression models that included summary risk estimates from non-AML and nonhuman studies as prior information. Our complete dataset included six human AML studies, three human leukemia studies, 10 human biomarker studies, and four experimental animal studies.

**Results:** A linear meta-regression model with intercept best predicted AML risks after cross-validation, both for the full dataset and AML studies only. Risk estimates in the low exposure range [<40 parts per million (ppm)-years] from this model were comparable, but more precise when the ERC was derived using all available data than when using AML data only. Allowing for between-study heterogeneity, RRs and 95% prediction intervals (95% PI) at 5 ppm-years were 1.58 (95% PI, 1.01–3.22) and 1.44 (95% PI, 0.85–3.42), respectively.

**Conclusions:** Integrating the available epidemiologic, biomarker, and animal data resulted in more precise risk estimates for benzene exposure and AML, although the large between-study heterogeneity hampers interpretation of these results. The harmonization steps required to fit the Bayesian meta-regression model involve a range of assumptions that need to be critically evaluated, as they seem crucial for successful implementation.

**Impact:** By describing a framework for data integration and explicitly describing the necessary data harmonization steps, we hope to enable risk assessors to better understand the advantages and assumptions underlying a data integration approach.

*See related commentary by Keil, p. 695*

## Introduction

There is international consensus that benzene exposure is causally related to acute myeloid leukemia (AML; ref. 1), but accurate description of the AML–benzene exposure-response curve (ERC) is still needed for impact and risk assessment. Vlaanderen and colleagues (2) used meta-regression to derive an ERC for total leukemia, which includes AML, but also subtypes for which a causal relation to benzene exposure has not been shown. Derivation of a precise ERC for AML was not considered possible because of the limited number of studies and low case numbers at low levels of exposure. Combining human AML data with data from closely related study domains, for example, human epidemiologic data on leukemia or cancer biomarkers, and animal experimental data could be used to increase precision of the estimated ERC in this, and other cases (3).

The aim of this article is to estimate the benzene-AML ERC using data from both human and animal studies and from both experimental and observational study designs. We also aim to identify and highlight some of the assumptions underlying the harmonization steps required for integrating data across these different study domains. We hypothesize that the shape of the benzene-AML ERC may be estimated more precisely when using a larger evidence base, but also that it will improve our ability to address concerns regarding the generalizability of results obtained from a small set of studies. We focus on the exposure-response relation at relatively low exposure levels, because the risks at low occupational and environmental benzene exposure levels are still being debated (4).

For this purpose, we add to the available epidemiologic studies that directly investigated benzene and AML: (i) human studies on benzene-induced leukemia, (ii) human biomarker studies in benzene-exposed workers on the induction of chromosomal aberrations (CA) and micronuclei (MN), and (iii) experimental animal studies on benzene-induced hematopoietic and lymphoid cancers. Biomarker studies for CA and MN were included because these were found to be associated with increased risk of cancer in large prospective cohort studies (5, 6).

We choose a Bayesian meta-regression approach because it can be readily adapted to include prior information on likely effect sizes and between-study heterogeneity (7), while also accounting more directly for the imprecisely estimated between-study heterogeneity than frequentist models (8).

## Materials and Methods

We refer to the Supplementary for a description of how individual studies were selected for the epidemiologic, human biomarker, and animal experimental study domains (Supplementary Study Selection).

[1] Institute for Risk Assessment Sciences, Utrecht University, Utrecht, the Netherlands. [2] The Netherlands Organisation for Applied Scientific Research (TNO), Zeist, the Netherlands.

**Note:** Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

B. Scholten, L. Portengen, J. Vlaanderen, and R. Vermeulen contributed equally to this article.

**Corresponding Author:** Bernice Scholten, Institute for Risk Assessment Sciences, Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, the Netherlands. Phone: 316-1537-2861; E-mail: bernice.scholten@tno.nl

Cancer Epidemiol Biomarkers Prev 2022;31:751–7

**doi:** 10.1158/1055-9965.EPI-21-0287

©2021 American Association for Cancer Research

A "reference" ERC for benzene and AML was estimated using a Bayesian version of the meta-regression model used by Vlaanderen and colleagues (4) for estimating the ERC for benzene and leukemia, and using only data from epidemiologic studies that directly investigated risk of AML in benzene-exposed workers. As usual for this type of (meta-regression) model, the input data consisted of reported summary risk estimates (i.e., log-HRs) and SEs for each (AML) study and for each (average) cumulative benzene exposure level for which it was reported. Additional details on the structure of the meta-regression model are provided further below.

To estimate an "augmented" ERC, i.e., an ERC that includes information from studies in some or all of the related study domains, we converted the available exposure-response information to fit the same format as that used for the AML studies, i.e., into log relative effect estimates with SEs at an estimated cumulative benzene exposure level. Detailed examples of how the available information was processed in order to fit this format (i.e., the data harmonization steps) are provided in the Supplementary Data Harmonization (including Supplementary Table S1), but the approach is also outlined below, separately for each study domain.

### Epidemiologic studies

For each study, risk estimates were selected from models where cumulative exposure was entered as a categorical variable, with SEs estimated from reported confidence intervals. No distinction was made between rate or HRs, standardized mortality ratios (SMR), or ORs, and we will refer to each of these using the term RR in the remainder of the article. When no cases were observed in one of the categories (as was the case for the study by Collins and colleagues (9), we imputed half a case as a continuity correction to allow calculation of the log$_{\mathrm{RR}}$ for the meta-regression. An overview of all included AML or leukemia studies can be found in the Supplementary AML studies (Supplementary Table S2) and Supplementary Leukemia studies (Supplementary Table S3), respectively.

Exposure estimates were based either on reported average cumulative exposures or, when these were not available, by assigning the midpoint of the reported range or, for open-ended upper categories, the lower category boundary multiplied by 5/3.

### Human biomarker studies

The meta-analytical approach of Scholten and colleagues (ref. 10; where benzene exposure effects are estimated on an additive scale) was modified in two ways to allow inclusion of the study data in our meta-regression framework. First, to quantify effects on a multiplicative scale, we log-transformed the reported proportions and used the delta method to estimate their variance on the log-scale (11). We then subtracted the log-transformed proportion of aberrant cells in the unexposed (or low-exposed) from that in the exposed and estimated its SE using standard variance rules. The main assumption underlying this approach is that the ratio of CA (or MN) in exposed to unexposed categories can be used to inform our prior estimate for the RR of benzene-induced AML. Second, to harmonize exposure levels in the biomarker studies to those in the epidemiologic studies, (average) exposure levels were multiplied by the reported average working histories, as a way to estimate cumulative exposures. If for a specific study no working history was available (which was the case for three CA and five MN studies) we assigned the average working history across all other studies. The main assumption underlying this step is that group differences in average exposures calculated from cross-sectional studies are equivalent to the differences observed in the epidemiologic studies.

### Animal experimental studies

Effect estimates for most animal studies were either available as, or could be easily expressed as, risk ratios. Concordance between experimental animal and human epidemiologic data for AML is supported by results presented in the recent International Agency for Research on Cancer (IARC) monograph on tumor site concordance and mechanisms of carcinogenesis (12). We estimated cumulative benzene exposure levels by multiplying reported exposure levels [parts per million (ppm)] by reported exposure durations without applying any further conversion factors. Results from multiple experiments reported in a single paper ($n = 2$) were considered separate studies. An overview of all animal studies can be found in the Supplementary file (Supplementary Table S4).

### Bayesian meta-regression model

Our meta-regression model is formulated as a hierarchical two-level random intercept and slope model. The (level 1) model for the observed (log) RRs (Y) is:

$$Y \sim MVN\left(E(Y), \Sigma_y\right)$$

$$E\left(Y_{ij}\right) = \delta_{0,i} + \sum_{k=1}^{K} \delta_{k,i} * X_{k,ij}$$

Where MVN is the multivariate normal distribution and E(Y) the expected value of Y, $i$ and $j$ index different studies and exposure levels, $\delta_{0,i}$ and $\delta_{k,i}$ are the study-specific intercept and slope coefficients, and where $X_k$ is either the benzene exposure level itself (for linear models; K = 1) or the $k^{\mathrm{th}}$ basis of a K-dimensional regression spline. The covariance matrix of residual errors ($\Sigma_y$) is assumed to be known and was estimated using the method proposed by Greenland and Longnecker (13).

The (level 2) model for the random intercept ($\delta_{0,i}$) and slope coefficient(s) ($\delta_{k,i}$) is:

$$\begin{bmatrix} \delta_{0,i} \\ \vdots \\ \delta_{\,k,i} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{\,k} \end{bmatrix}, \sum_{\delta}\right)$$

We aimed to specify priors that are strong enough to rule out unreasonable values, while still allowing the data to dominate the prior when it provides enough information.

We assigned a weakly informative normal prior centered at 0 and with a scale of 1 to the overall intercept ($\beta_0$), and normal priors centered at 0 with a scale of 2 to the slope parameter(s) ($\beta_1 \cdots \beta_K$). This reflects our prior belief that the RR at zero exposure is likely to be exp (0) = 1 and rather unlikely (i.e., with prior probability <20%) to be outside the range exp(±1.28) = [1/3.6, 3.6] and that the RR per 100 ppm-years is unlikely to be outside the range exp($\pm 2^{*}1.28$) = [1/13, 13] (for the linear model).

For the random effect variances we followed the recommendations by Röver and colleagues (14) in using a half-Cauchy prior for the random effect variances and the prior suggested by Lewandowski, Kurowicka, and Joe (LKJ; ref. 15) for the correlations between random effects. Our main analyses are based on using a half-Cauchy with a scale of 1. Our choice for this particular prior scale reflects our belief that the between-study variation in log$_{\mathrm{RRs}}$ is likely (i.e., with prior probability >80%) within the range (0.16–6.3), but with a mode at 0, and it allows for considerable heterogeneity in RRs between studies. Based on this prior, and assuming an overall RR of 1, approximately 20% of study-specific RRs (per 100 ppm-years) are expected to be

Downloaded from http://aacrjournals.org/cebp/article-pdf/31/4/751/3116574/751.pdf by guest on 26 September 2023

between 0.9 and 1.1, 40% are expected to be between 0.7 and 1.5, 60% are expected to be between 0.4 and 2.4, and 80% are expected to be between and 0.1 and 9.9.

The parameter η of the LKJ prior was set at 2, which mainly serves to exclude very strong correlations between estimated random effect parameters [e.g., 80% of the intercept-slope correlations is expected to be in the range (−0.6 to +0.6)].

### Exposure–response models

To assess the shape of the ERC, we fitted a regression spline model using a natural regression spline basis with interior knots at 10 and 65 ppm-years (the approximate 33% and 67% percentiles of the exposure distribution) and outer knots at 0 and 130 ppm-years (the approximate 0% and 85% percentiles of the exposure distribution). Alternative exposure-response models included a regression spline model without an intercept, forcing the ERC through the origin (RR = 1) at zero exposure, and linear models with and without an intercept.

Between-study heterogeneity was accommodated by allowing for study-specific intercepts and slopes (regression coefficients) as (correlated) random effects. We additionally present 95% prediction intervals (95% PI) that take into account the between-study heterogeneity as recommended by Higgins and colleagues (16). Prediction intervals were calculated using estimates for the between-study heterogeneity for AML studies.

We evaluated model fit for each individual study by jackknifing (i.e., leaving out one study at a time), refitting the meta-regression model, and calculating the ratio of the sum of the differences between observed and predicted values over its estimated SE for each held out study as an externally studentized residual. An absolute value of the ratio exceeding 3 was considered evidence of severe lack of fit. To compare the quality of posterior predictions from different model structures we estimated the sum of the expected log pointwise predictive density (ELPD; ref. 17) using the models fitted during jackknifing. Higher ELPDs indicate better predictions for the left-out studies and can be calculated for the full set of studies, but also for a subset (e.g., only studies in the AML set).

Finally, to combine predictions from models with different model structures (i.e., linear/spline, with/without intercept) we used Bayesian stacking (18), with weights calculated based on the ELPD estimates. We call the resulting model the “consensus” model and the estimated ERC the “consensus” ERC.

### Sensitivity analyses

First, we evaluated the sensitivity of our results to our choice of prior for the between-study covariance by changing the scale of the half-Cauchy prior to 0.5 and 5.

Second, we investigated the effect of excluding risk estimates for relatively high exposures (i.e., over 40 ppm-years) from the model on estimated risks at low exposures.

Third, we used an approach similar to that described in Bartell and colleagues (19), reducing the precision of point estimates for a subset of studies, to illustrate how our approach could be used to selectively downweigh the impact of some studies, in this case adding an (arbitrary) 10-fold uncertainty factor to results from the experimental animal studies and a three-fold uncertainty factor to results from both sets of biomarker studies.

### Impact assessment

To illustrate how our framework could be used for risk assessment, we calculated the excess risk of AML due to benzene exposure for our “consensus” ERC using a life table analysis. Background incidence rates for AML in the Netherlands were obtained from the Dutch Integraal Kanker Centrum (IKC) and combined with Dutch mortality rates as obtained from Statistics Netherlands. Excess risk of AML was estimated for workers that were exposed to 0.1 ppm benzene for 40 years (between age 20–60), assuming a 5-year lag period, and evaluated at age 80. We also estimated benzene exposure levels corresponding to the definition of acceptable risk (AR) and maximum tolerable risk (MTR) levels for occupational settings, i.e., the benzene exposure level at which the number of excess cases is either 40 per 1,000,000 or 40 per 10,000 exposed workers.

### Software

The Bayesian meta-regression model was implemented in STAN using the *brms* package (version 2.10.0). We collected 10,000 samples for each parameter from 4 chains after a burn-in of 5,000 iterations using Markov Chain Monte Carlo (MCMC) techniques to sample from the posterior distribution. The algorithm was tuned by increasing adapt.delta to 0.999 to avoid divergent transitions (20). The posterior distribution was summarized by calculating the mean, SD, and 2.5, 50, and 97.5 percentiles.

Jackknifing (leave-one-group-out cross-validation) was performed using the R function *kfold* from the *brms* package, which was also used to estimate ELPDs for the held-out studies. Weights for Bayesian model averaging (model stacking) used to estimate the “consensus” ERC were estimated using the *stacking_weights* function from the *loo* package in R.

## Results

We removed the MN biomarker study by Surrallés and colleagues (21) from all further analyses because the results of the (first) jackknifing analysis indicated a severe lack of fit for this study even in our most flexible exposure-response model (i.e., a regression spline model with intercept; Supplementary Table S5). Our final study base therefore consisted of 26 studies: six human epidemiologic studies on AML, three human leukemia studies (that had not reported on AML), 10 biomarker studies (4 CA and 6 MN), and seven experimental animal studies (from four publications; refs. 22–25). Most of the AML and leukemia studies, but only a minority of other studies, had estimates for more than one exposure level.

**Table 1** shows estimated RRs at cumulative benzene exposure levels of 0, 5, 10, 20, and 40 ppm-years based on our spline model with intercept for models fitted to data from studies in the reference set (i.e., human AML studies), after adding studies from single additional study domains, and after using all studies from all domains. Using data from the full set of studies resulted in slightly higher relative risk estimates than using the reference set only, and these were estimated more precisely with narrower confidence intervals (CI) and PI. There was considerable between-study heterogeneity in risk estimates however, as evident also from the PIs that were much wider than the CIs. Using the full set of studies resulted in narrower PIs than using the reference set only, suggesting that the gain in precision for population-level risk estimates was not at the expense of increased between-study heterogeneity.

Using cross-validation to estimate and compare the quality of model predictions, we found that the linear model with intercept provided the best predictions (i.e., had lowest kICs; see **Table 2**); both for the full set of studies and when considering predictions only for the AML studies. This was true also when models were fitted using data only from AML studies. Predictions for AML studies were also better from models fitted using all available data than from models fitted to the AML data

Downloaded from http://aacrjournals.org/cebp/article-pdf/31/4/751/3116574/751.pdf by guest on 26 September 2023

**Table 1.** Relative risk estimates and 95% CIs and PIs for benzene-induced AML, for benzene exposure at selected exposure levels.

| Benzene (ppm, y) | AML | AML + leukemia | AML + CA biomarker | AML + MN biomarker | AML + animal data | All |
|---|---|---|---|---|---|---|
| 0[a] | 1.39 | 1.40 | 1.35 | 1.59 | 1.44 | 1.58 |
| | (0.77–2.47) | (0.81–2.41) | (0.76–2.34) | (1.15–2.12) | (0.82–2.44) | (1.17–2.09) |
| | (0.73–3.67) | (0.76–3.54) | (0.73–3.45) | (0.94–3.53) | (0.78–3.55) | (0.96–3.47) |
| 5 | 1.40 | 1.36 | 1.41 | 1.60 | 1.35 | 1.57 |
| | (0.84–2.30) | (0.87–2.09) | (0.91–2.18) | (1.26–1.99) | (0.86–2.07) | (1.25–1.92) |
| | (0.72–3.89) | (0.73–3.49) | (0.78–3.68) | (0.92–3.69) | (0.73–3.39) | (0.93–3.51) |
| 10 | 1.40 | 1.32 | 1.48 | 1.62 | 1.28 | 1.57 |
| | (0.79–2.50) | (0.84–2.08) | (0.98–2.19) | (1.24–2.09) | (0.81–2.03) | (1.25–1.90) |
| | (0.60–4.83) | (0.61–4.05) | (0.71–4.48) | (0.78–4.40) | (0.58–3.80) | (0.81–3.98) |
| 20 | 1.44 | 1.32 | 1.61 | 1.69 | 1.23 | 1.61 |
| | (0.66–3.32) | (0.72–2.45) | (0.95–2.52) | (1.13–2.45) | (0.68–2.28) | (1.18–2.05) |
| | (0.43–7.60) | (0.45–5.63) | (0.57–6.73) | (0.60–6.18) | (0.41–5.19) | (0.65–5.28) |
| 40 | 1.64 | 1.56 | 1.79 | 1.94 | 1.37 | 1.78 |
| | (0.70–4.08) | (0.81–3.05) | (1.05–2.86) | (1.20–2.97) | (0.75–2.63) | (1.28–2.34) |
| | (0.43–10.21) | (0.47–7.57) | (0.57–8.55) | (0.60–8.12) | (0.40–6.73) | (0.63–6.52) |

Note: Risk estimates are population-level estimates from a meta-regression model that includes an intercept and allows for nonlinear effects of benzene exposure. Estimates are given only for AML studies, and with inclusion of several data domains, including the full dataset. Second row in each cell, CI; third row, PI.

[a] Intercept.

only. As a result, estimated model weights were highly skewed towards the linear model with intercept, which was assigned a weight of nearly 1 (versus <1e-5 for any of the other model structures), making the "consensus" model virtually identical to the linear model with intercept.

Estimated ERCs using the linear "consensus" model for the AML set only and after adding studies from different domains are presented graphically in **Fig. 1**. Although most datapoints from the biomarker and experimental animal studies were in the lower exposure region, inclusion of these studies had only limited effect on the estimated ERC. Similar to the more flexible regression spline model, estimated RRs at low cumulative benzene exposure levels were generally lower and less precise when using data only from the AML studies, when compared with using all available studies (**Table 3**). As an example, the risk estimate (95% PI) at 5 ppm-years was 1.44 (95% PI, 0.85–3.42) when using only AML studies, while it was 1.58 (95% PI, 1.01–3.22) when using all available studies.

Parameter estimates for the consensus parameters and between-study (co)variance for the linear model with an intercept fitted to the full dataset are provided in the Supplemental Materials (Supplementary Table S6). Between-study variance in slopes was large relative to the consensus slope estimate. There were no large differences in estimated between-study variance for slopes for different study domains, but these were estimated rather imprecisely.

## Sensitivity analyses

### Prior choice for the between-study (co)variance

Detailed results from sensitivity analyses regarding our prior choice for the between-study (co)variance are presented in the Supplementary file (Supplementary Tables S7-S9).Using a more diffuse prior had relatively little effect on the width of CIs for predicted exposure effects, but resulted in significantly wider posterior PIs. As an example, for the linear model with intercept the CIs for the predicted exposure effect at 40 ppm-yrs were (1.58–2.04) and (1.51–2.11) for prior scales of 0.5 and 5 respectively, while the corresponding PIs were (1.03–3.72) and (0.80–5.54).

Prior-posterior plots (Supplementary Fig. S1) suggest that there is no large mismatch between any of these priors and information provided by the data itself.

### Excluding high exposure datapoints

Risk estimates were comparable but less precise after excluding all datapoints with benzene exposures over 40 ppm-years, with a noticeable drop in risk estimates for exposures of 40 ppm-years for the regression-spline based models (Supplementary Table S10).

### Uncertainty factors

Estimated risks were marginally lower and considerably less precise when three-fold uncertainty factors were used for the biomarker

**Table 2.** Comparison of predictive model quality for models with different model structures.

| Model | Full dataset | Full dataset : AML only | AML |
|---|---|---|---|
| Linear model without intercept | 121.6 | 54.52 | 56.82 |
| Linear model with intercept | 85 | 49.12 | 52.76 |
| Spline model without intercept | 113.2 | 54.02 | 58.88 |
| Spline model with intercept | 96 | 52.16 | 56.56 |

Note: Results are presented as the sum of ELPD ($\sum ELPD$), with higher values indicating better model predictions. ELPDs were estimated using jackknifing (leave-one-study-out cross-validation). Results are shown for models fitted to the full set of studies from all study domains for all studies and for AML studies only and for models fitted to data only from AML studies.

Downloaded from http://aacrjournals.org/cebp/article-pdf/31/4/751/3116574/751.pdf by guest on 26 September 2023

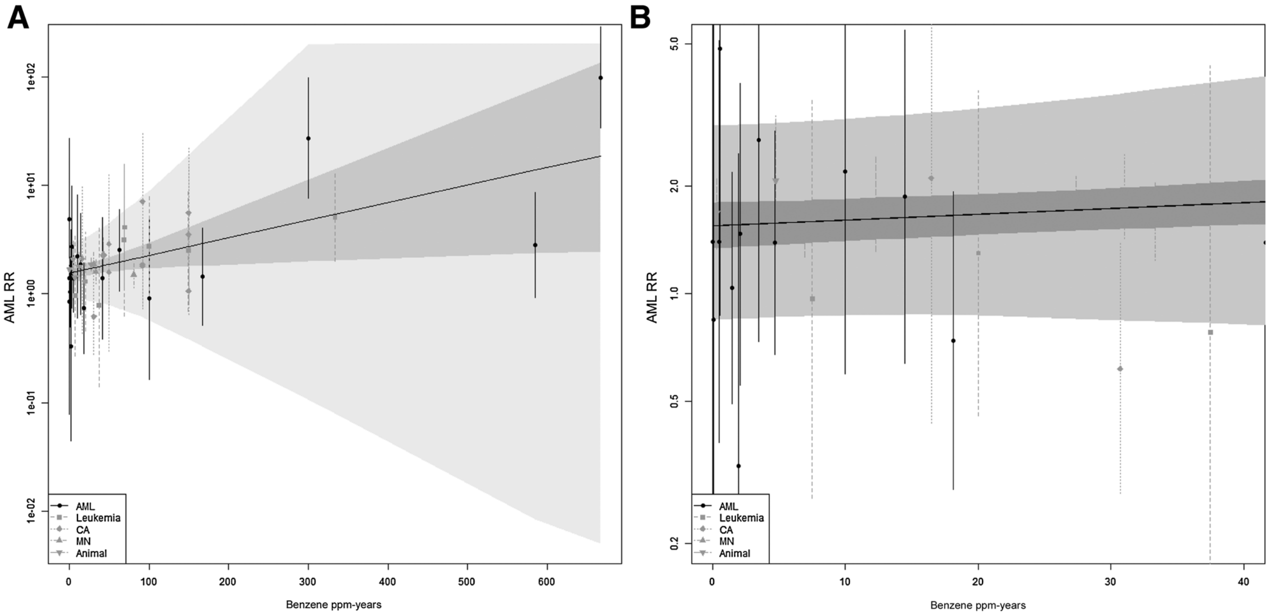

**Figure 1.**

Benzene exposure and RR of developing AML based on various data domains. Linear model with intercept. Dark ribbon, confidence interval for all studies; light ribbon, PI for all studies. Dashed line represents knots. **A,** all study points; **B,** zoomed 0 to 40 ppm-years.

studies and a 10-fold uncertainty factor for the experimental animal studies (Supplementary Table S11). The effect was more pronounced for CIs: risk estimates at 5 ppm-years were reduced from 1.58 (95% CI, 1.37–1.82) to 1.52 (95% CI, 1.16–2.00) when uncertainty factors were applied.

### Impact assessment

Results from the life table analyses that were used to estimate the excess risk of AML due to benzene exposure using the linear model exposure-response model with intercept are presented in **Table 4**. These results indicate that, when workers are exposed to 0.1 ppm for 40 years, the estimated number of excess cases is higher than that corresponding to AR levels, except when the most conservative approach (i.e., subtracting the intercept) is used on the linear model fitted to data from only the reference set, where it is just slightly lower.

Exposure levels corresponding to AR and the MTR levels are shown in the Supplementary Tables S12 and S13, and range from 0.0003 ppm (for the interpolated linear model) to 0.098 ppm (for the model with intercept subtracted) when models are fitted to the full set of studies.

## Discussion

To estimate the benzene-AML ERC, we collected and summarized human and animal studies. The ERC derived using studies

**Table 3.** Risk estimates and 95% PIs for benzene-induced AML at selected exposure levels.

| Benzene exposure (ppm, y) | AML | AML + leukemia | AML + CA biomarker | AML + MN biomarker | AML + animal data | All |
|---|---|---|---|---|---|---|
| 0<br>(intercept) | 1.42<br>(0.91–2.21)<br>(0.84–3.40) | 1.42<br>(0.99–2.06)<br>(0.86–3.22) | 1.54<br>(1.11–2.12)<br>(0.94–3.28) | 1.56<br>(1.31–1.90)<br>(1.01–3.18) | 1.33<br>(0.91–1.93)<br>(0.80–3.03) | 1.55<br>(1.34–1.80)<br>(1.00–3.16) |
| 5 | 1.44<br>(0.94–2.22)<br>(0.85–3.42) | 1.45<br>(1.01–2.09)<br>(0.88–3.29) | 1.56<br>(1.14–2.11)<br>(0.97–3.32) | 1.60<br>(1.36–1.92)<br>(1.03–3.27) | 1.36<br>(0.93–1.96)<br>(0.82–3.09) | 1.58<br>(1.37–1.82)<br>(1.01–3.22) |
| 10 | 1.46<br>(0.95–2.24)<br>(0.85–3.50) | 1.48<br>(1.04–2.12)<br>(0.89–3.35) | 1.58<br>(1.17–2.10)<br>(0.97–3.38) | 1.63<br>(1.39–1.94)<br>(1.04–3.35) | 1.38<br>(0.96–1.98)<br>(0.83–3.17) | 1.61<br>(1.40–1.84)<br>(1.02–3.30) |
| 20 | 1.50<br>(0.97–2.31)<br>(0.81–3.77) | 1.54<br>(1.08–2.19)<br>(0.88–3.58) | 1.63<br>(1.23–2.10)<br>(0.95–3.58) | 1.71<br>(1.43–2.03)<br>(1.02–3.68) | 1.44<br>(1.01–2.04)<br>(0.83–3.40) | 1.67<br>(1.46–1.90)<br>(1.02–3.51) |
| 40 | 1.59<br>(0.96–2.55)<br>(0.66–4.72) | 1.67<br>(1.16–2.38)<br>(0.81–4.39) | 1.72<br>(1.29–2.24)<br>(0.82–4.37) | 1.86<br>(1.44–2.37)<br>(0.90–4.86) | 1.55<br>(1.10–2.21)<br>(0.76–4.21) | 1.80<br>(1.56–2.07)<br>(0.94–4.21) |

Note: Risk estimates are population-level estimates from a linear meta-regression model that includes an intercept. Estimates are given for only AML studies, and with inclusion of several data domains, including the full dataset. Second row in each cell, CI; third row: PI.

Downloaded from http://aacrjournals.org/cebp/article-pdf/31/4/751/3116574/751.pdf by guest on 26 September 2023

**Table 4.** Results of a life table analysis for calculating excess cases per 1,000,000 at age 80 when exposed to benzene at 0.1 ppm for 40 years.

| | AML | AML + leukemia | AML+ all data | <40 ppm-yrs |
|---|---|---|---|---|
| Linear with intercept | 1,289 | 1,322 | 1,698 | 1,711 |
| Linear with intercept – intercept subtracted | 30 | 42 | 41 | 3 |
| Linear with intercept – interpolation | 1,222 | 1,253 | 1,606 | 1,619 |
| Spline no intercept | 295 | 204 | 379 | — |
| Linear no intercept | 64 | 65 | 75 | 184 |

from this broader evidence base was very similar to that based on data from the human AML studies only, but with more precise risk estimates. Based on results from cross-validation, prediction of risks observed in single (held-out) AML studies was improved by using data from other study domains.

We included a detailed description of data harmonization steps and our prior motivation to improve understanding of the assumptions underlying the use of these models by risk assessors. The need to harmonize exposure and outcome variables required us to make strong and mostly untestable assumptions. For human studies, exposure metrics used in the, mostly cross-sectional, biomarker studies (average exposure levels) were fundamentally different from those in the long-term prospective epidemiologic studies (cumulative exposure levels). In addition, effects recorded in the biomarker studies had to be converted to relative effects to allow combination with epidemiologic data. For the animal data, cumulative exposure could be readily estimated, but it is unclear whether any interspecies extrapolation factors should have been applied. Although others (e.g., Bartell and colleagues; ref. 19) used conversion factors, there seems to be no general agreement on either use or value. We chose not to apply any conversion factor, based on the argument that metabolic rate and cell division are roughly inversely correlated to lifespan (26).

Uncertainties stemming from the extrapolation of risks from animal to human studies in current risk assessment procedures are typically addressed by using a fixed set of interspecies extrapolation factors (27). In an analogous approach, we (arbitrarily) used a 10-fold uncertainty factor for the animal studies and a three-fold uncertainty factor for the biomarker studies to downweigh evidence from these study domains, resulting in slightly less precise risk estimates. While safeguarding against inappropriate over-reliance on animal or biomarker data, this may also result in suboptimal use of the available data.

Prior choice for the between study variation is an important ingredient of Bayesian meta-analyses, and may be crucial when there is little information. There were only few studies per study domain in the analysis, and we therefore used a half-Cauchy prior with a scale parameter of 1, to exclude implausible high values for the random effects variance. We evaluated the sensitivity of our findings to this choice in sensitivity analyses (Supplementary Table S7; Supplementary Figs. S1 and S2), with the results confirming our prior is broadly compatible with the data, but also that estimated heterogeneity, which affects precision of e.g., common slope factors, is quite sensitive to the prior scale. We discuss the importance of heterogeneity for model inference further below, but note that with more or more precise data, the meta-regression model could be formulated with hyperpriors for the half-Cauchy scale. Our approach also allows evaluation and comparison of study heterogeneity across different study domains. We found no large differences in heterogeneity between study domains in our study, this could well be due to the fact that these were estimated rather poorly, as there were only few studies per study domain.

We used life table calculation to evaluate the impact of our estimated "consensus" ERC for further risk assessment. It should be noted that, in the presence of significant between-study heterogeneity, use of an ERC based on consensus (population-level) parameters may be difficult to justify (28). Without knowing, or at least suspecting, what the reasons for the apparent heterogeneity may be, the mean parameters may be difficult to interpret. In case of strong heterogeneity, the mean parameters are also estimated with (near) equal weights for smaller and larger studies, which could be problematic when smaller studies are more likely to suffer from small-sample or publication bias (29). Alternatively, the full (random effects) distribution may be used e.g., to average lifetable results across a random sample of study-specific ERCs or for choosing an upper quantile of the between-study distribution under the assumption that the higher risks are observed in better studies [similar to what is done in benchmark dose modeling (BMDL)].

Sobel and colleagues (30) and Dahabreh and colleagues (28) recently discussed the problem of casual interpretation of results from meta-analyses. Both papers stress the importance of investigating sources of heterogeneity and rely on using additional individual-level covariate data to account for differential selection and exposure effects. Our approach is flexible enough to include further covariates as moderators.

We found that a linear model with intercept provided the best predictions. Possible reasons for this intercept include exposure measurement error, uncontrolled confounding, and healthy worker selection. Models with an intercept present considerable interpretational difficulties from a risk assessment perspective, and risk assessors therefore often prefer (meta-regression) models that do not allow intercepts. We therefore evaluated risks using a number of different approach to account for the intercept and also assessed risks based on the regression-spline model without an intercept. Estimated AR levels from these models estimated using all avalilable data were in the range of 0.0003 to 0.098 ppm, which includes the limit for benzene exposure that was recently proposed by the Risk Assessment Comittee of 0.05 ppm (31).

Our approach differs from earlier proposals for using Bayesian methods to integrate data for risk assessment (e.g., Bartell and colleagues, Dumouchel and colleagues; ref. 19, 32). Most notably these authors used the additional data to derive a single prior for a slope coefficient from a linear meta-analytical model. In contrast, we aimed to include data from several different study domains simultaneously and wanted to allow for potential nonlinear effects using a meta-regression model. Our approach is more easy to use with more complex model structures and allows more explicit evaluation of between-study heterogeneity for studies from different study sources.

Our approach can be seen as a first step towards quantitative integration of human and animal data in risk assessment. It involves a tradeoff between a potential gain in statistical precision that comes with a more complete evaluation of the evidence base, but it also runs the risk of introducing bias or increasing heterogeneity due to the harmonization steps that may be required. Risk assessors may therefore choose to use this approach primarily when evaluating compounds with weak epidemiologic evidence (i.e., insufficient data) but a

Downloaded from http://aacrjournals.org/cebp/article-pdf/31/4/751/3116574/751.pdf by guest on 26 September 2023

large animal or molecular evidence base. Implementation of our approach in regulatory chemical risk assessment exercises would require additional knowledge (such as better understanding of toxicokinetics, better evidence for prediagnostic biomarker-cancer associations, better understanding of reasonable approaches to extrapolate animal evidence to the human setting) to reduce the assumptions that had to be made in the current evaluation.

To conclude, we provide a first step towards the quantitative integration of data from different study domains, into human risk assessment and identified a number of gaps that need to be addressed in further research.

## Authors' Disclosures

B. Scholten reports grants from Institute for Risk Assessment Sciences (IRAS), Utrecht University, and The Netherlands Organisation for Applied Scientific Research (TNO) during the conduct of the study. R. Stierum reports grants from IRAS, Utrecht University, and TNO during the conduct of the study. No disclosures were reported by the other authors.

## Authors' Contributions

**B. Scholten:** Conceptualization, resources, formal analysis, investigation, writing–original draft. **L. Portengen:** Software, supervision, validation, investigation, writing–original draft. **A. Pronk:** Conceptualization, methodology. **R. Stierum:** Methodology, writing–original draft. **G.S. Downward:** Writing–original draft. **J. Vlaanderen:** Conceptualization, writing–original draft. **R. Vermeulen:** Conceptualization, resources, supervision, writing–original draft.

## Acknowledgments

This work was supported by internal funds from both IRAS and TNO. We thank three anonymous reviewers for their helpful suggestions that greatly improved the paper.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received March 2, 2021; revised June 21, 2021; accepted November 22, 2021; published first December 13, 2021.

Downloaded from http://aacrjournals.org/cebp/article-pdf/31/4/751/3116574/751.pdf by guest on 26 September 2023

## References

1. IARC. Benzene. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans Volume 120. Lyon, France: IARC; 2018.
2. Vlaanderen J, Rothman N, Portengen L, Lan Q, Kromhout H, Vermeulen R. Flexible meta-regression to assess the shape of the benzene–leukemia exposure–response curve. Environ Health Perspect 2010;118:526–32.
3. National Academies of Sciences Engineering, Medicine, Division on Earth and Life Studies, Board on Environmental Studies and Toxicology, Committee on Incorporating 21st Century Science into Risk-Based Evaluations. Using 21st century science to improve risk-related evaluations. Washington (DC): National Academies Press; 2017.
4. Mchale CM, Zhang L, Smith MT. Current understanding of the mechanism of benzene-induced leukemia in humans: implications for risk assessment. Carcinogenesis 2012;33:240–52.
5. Bonassi S, Norppa H, Ceppi M, Strömberg U, Vermeulen R, Znaor A, et al. Chromosomal aberration frequency in lymphocytes predicts the risk of cancer: results from a pooled cohort study of 22 358 subjects in 11 countries. Carcinogenesis 2008;9:1178–83.
6. Bonassi S, Znaor A, Ceppi M, Lando C, Chang WP, Holland N, et al. An increased micronucleus frequency in peripheral blood lymphocytes predicts the risk of cancer in humans. Carcinogenesis 2007;28:625–31.
7. Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JPT. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. Stat Med 2015;34:984–98.
8. Jackson D, Bowden J, Baker R. How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? J Stat Plan Inference 2010;140:961–70.
9. Collins JJ, Anteau SE, Swaen GMH, Bodner KM, Bodnar CM. Lymphatic and hematopoietic cancers among benzene-exposed workers. J Occup Environ Med 2015;57:159–63.
10. Scholten B, Vlaanderen J, Stierum R, Portengen L, Rothman N, Lan Q, et al. A quantitative meta-analysis of the relation between occupational benzene exposure and biomarkers of cytogenetic damage. Environ Health Perspect 2020;128: 87004.
11. Oehlert GW. A note on the delta method. Am Stat 1992;46:27–9.
12. IARC. Tumour site concordance and mechanisms of carcinogenesis. IARC Scientific Publication No. 165. Baan RA, Stewart BW, Straif K, editors. Lyon, France: IARC; 2019.
13. Greenland S, Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. Am J Epidemiol 1992; 135:1301–9.
14. Röver C, Bender R, Dias S, Schmid CH, Schmidli H, Sturtz S, et al. On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. Res Synth Methods 2021;12:448–74.
15. Lewandowski D, Kurowicka D, Joe H. Generating random correlation matrices based on vines and extended onion method. J Multivar Anal 2009;100:1989–2001.
16. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J Roy Stat Soc Ser A Stat Soc 2009;172:137–59.
17. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput 2017;27:1413–32.
18. Yao Y, Vehtari A, Simpson D, Gelman A. Using stacking to average bayesian predictive distributions (with Discussion). Bayesian Anal 2018;13:917–1007.
19. Bartell SM, Hamra GB, Steenland K. Bayesian analysis of silica exposure and lung cancer using human and animal studies. Epidemiology 2017;28:281–7.
20. Stan Development Team. Brief guide to Stan's warnings. 2019. Available from: https://mc-stan.org/misc/warnings.html.
21. Surrallés J, Autio K, Nylund L, Järventaus H, Norppa H, Veidebaum T, et al. Molecular cytogenetic analysis of buccal cells and lymphocytes from benzene-exposed workers. Carcinogenesis 1997;18:817–23.
22. Cronkite EP, Drew RT, Inoue T, Hirabayashi Y, Bullis JE. Hematotoxicity and carcinogenicity of inhaled benzene. Environ Health Perspect 1989;82: 97–108.
23. Kawasaki Y, Hirabayashi Y, Kaneko T, Kanno J, Kodama Y, Matsushima Y, et al. Benzene-induced hematopoietic neoplasms including myeloid leukemia in Trp 53-deficient C57BL/6 and C3H/He mice. Toxicol Sci 2009;110:293–306.
24. Farris GM, Robinson SN, Gaido KW, Wong BA, Wong VA, Leonard L, et al. Effects of low concentrations of benzene on mouse hematopoietic cells in vivo: a preliminary report. Environ Health Perspect 1996;104:1275–6.
25. Li GX, Hirabayashi Y, Yoon BI, Kawasaki Y, Tsuboi I, Kodama Y, et al. Thioredoxin overexpression in mice, model of attenuation of oxidative stress, prevents benzene-induced hemato-lymphoid toxicity and thymic lymphoma. Exp Hematol 2006;34:1687–97.
26. Dutta S, Sengupta P. Men and mice: relating their ages. Life Sci 2016;152:244–8.
27. Jones DR, Peters JL, Rushton L, Sutton AJ, Abrams KR. Interspecies extrapolation in environmental exposure standard setting: a Bayesian synthesis approach. Regul Toxicol Pharmacol 2009;53:217–25.
28. Dahabreh IJ, Petito LC, Robertson SE, Hernán MA, Steingrimsson JA. Toward causally interpretable meta-analysis: transporting inferences from multiple randomized trials to a new target population. Epidemiology 2020; 31:334–44.
29. Richardson D, Cole SR, Ross R, Poole C, Chu H, Keil A. Meta-analysis and sparse-data bias. Am J Epidemiol 2020;190:336–40.
30. Sobel M, Madigan D, Wang W. Causal inference for meta-analysis and multi-level data structures, with application to randomized studies of Vioxx. Psychometrika 2017;82:459–74.
31. Committee for Risk Assessment. Opinion on scientific evaluation of occupational exposure limits for Benzene. Helsinki (Finland): ECHA; 2018. Available from: https://echa.europa.eu/documents/10162/13641/benzene_opinion_en.pdf/4fec9aac-9ed5-2aae-7b70-5226705358c7.
32. Dumouchel WH, Harris JE. Bayes methods for combining the results of cancer studies in humans and other species. J Am Stat Assoc 1983;78:293–308.