

RESEARCH

Open Access

# Carcinogen metabolism, cigarette smoking, and breast cancer risk: a Bayes model averaging approach

Nadine Stephenson, Lars Beckmann, Jenny Chang-Claude\*

## Abstract

**Background:** Standard logistic regression with or without stepwise selection has the disadvantage of not incorporating model uncertainty and the dependency of estimates on the underlying model into the final inference. We explore the use of a Bayes Model Averaging approach as an alternative to analyze the influence of genetic variants, environmental effects and their interactions on disease.

**Methods:** Logistic regression with and without stepwise selection and Bayes Model Averaging were applied to a population-based case-control study exploring the association of genetic variants in tobacco smoke-related carcinogen pathways with breast cancer.

**Results:** Both regression and Bayes Model Averaging highlighted a significant effect of *NAT1*\*10 on breast cancer, while regression analysis also suggested a significant effect for packyears and for the interaction of packyears and *NAT2*.

**Conclusions:** Bayes Model Averaging allows incorporation of model uncertainty, helps reduce dimensionality and avoids the problem of multiple comparisons. It can be used to incorporate biological information, such as pathway data, into the analysis. As with all Bayesian analysis methods, careful consideration must be given to prior specification.

## Background

Logistic regression and regression with stepwise selection are standard approaches to assess individual and joint effects of genetic and environmental factors on disease risk. However, one drawback is that the resulting estimates depend on the choice of the underlying causal model, and that hence a different set of covariates may lead to different effect estimates and potentially a different pattern of significance. Moreover, standard regression approaches do not incorporate the uncertainty about our choice of the assumed causal model into the final inference.

An alternative approach to analyze such data in combination is Bayes Model Averaging (BMA) [1], which explicitly accounts for uncertainty with respect to the causal model. BMA specifies prior distributions for model parameters and uses Markov Chain Monte Carlo

(MCMC) methods to infer posterior estimates from the priors and from the data. Its inherent model selection feature evaluates different submodels and inference is obtained by averaging over all models considered. By selecting and evaluating a range of submodels, BMA provides a means to reduce dimensionality in the presence of many predictors, when including all variables and their pairwise or higher-order interactions into a logistic model might lead to unstable estimates and bias due to sparse data and correlation [2]. Model selection methods like stepwise regression achieve a similar goal, but do so in a mechanical way, often leading to globally suboptimal and unstable estimates.

We applied both BMA and logistic regression with and without stepwise selection to data from a case-control study exploring the association of genetic variants in the cigarette smoke carcinogen metabolism and breast cancer. Cigarette smoke is known to contain aromatic amines and polycyclic aromatic hydrocarbons, whose conversion to reactive metabolites by catalyzing

\* Correspondence: j.chang-claude@dkfz-heidelberg.de  
Division of Cancer Epidemiology, German Cancer Research Center DKFZ, Im Neuenheimer Feld 581, 69120 Heidelberg, Germany

enzymes can lead to DNA damage as a first step in breast carcinogenesis. The present population-based case-control study of breast cancer in Germany evaluated the role of genetic polymorphisms in Phase I and II enzymes *NAT1* and *NAT2* in the AA pathway and *CYP1A1*, *CYP1B1*, *GSTM1* and *GSTT1* in the PAH pathway and cigarette smoke exposure in breast carcinogenesis. We analyzed pairwise interactions of polymorphisms as well as interactions of smoking and the polymorphisms to determine an effect on breast cancer risk. The postulated pathways are depicted as a directed acyclic graph in Figure 1.

## Methods

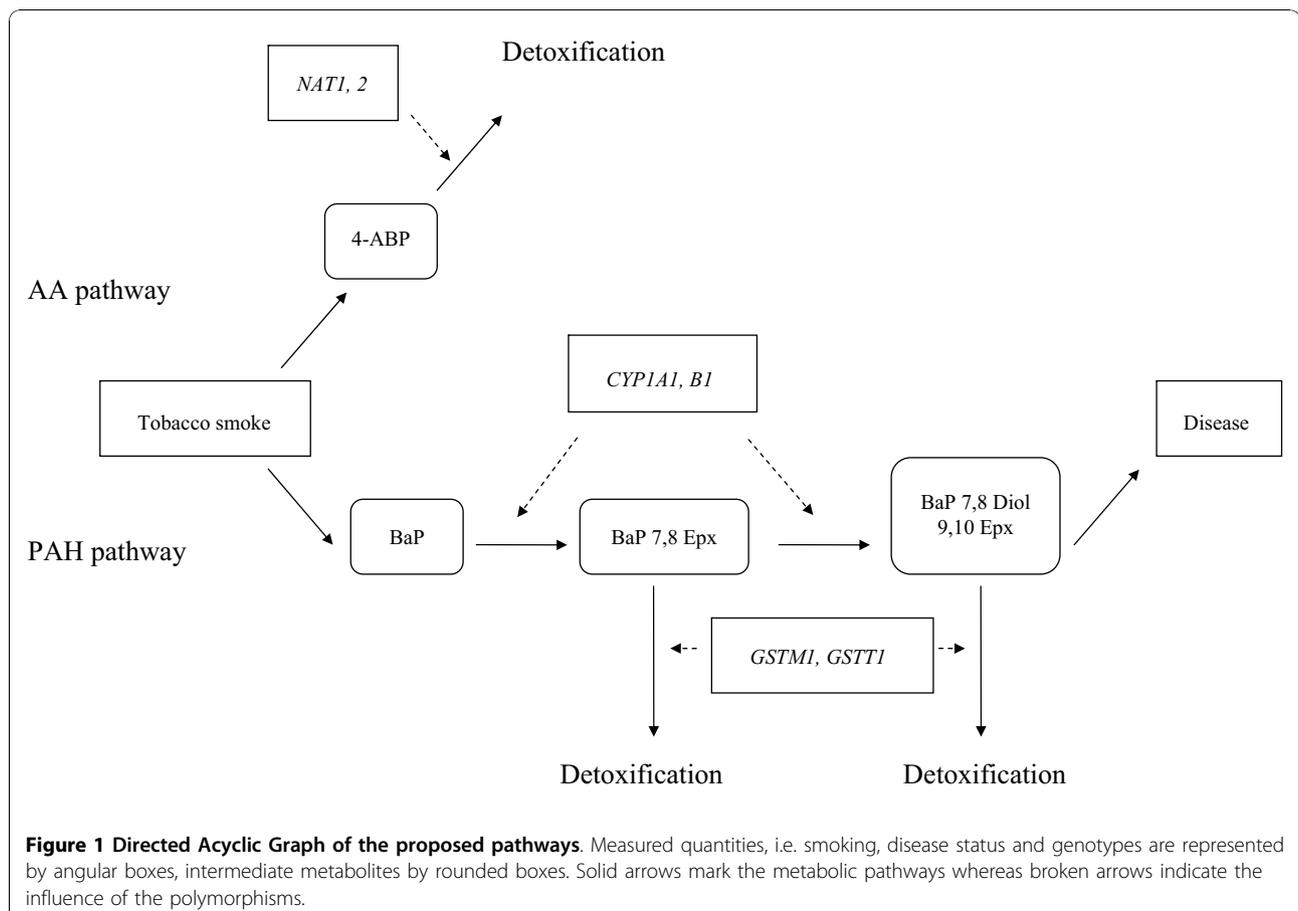
### Data

Data were derived from a population-based matched case-control study on breast cancer conducted in the study regions "Rhein-Neckar-Odenwald" and "Freiburg" of Southern Germany between 1992 and 1995, as described previously [3,4]. Cases were diagnosed by age 50 with invasive or in-situ breast cancer, and two controls were matched to cases by age at diagnosis and study region. Participants completed a self-administered questionnaire assessing demographic factors,

anthropometric measures and other known or putative risk factors, including smoking history. All study participants gave written informed consent, and the study was reviewed and approved by the ethics committee of the University of Heidelberg, Heidelberg, Germany.

Smoking behavior was assessed over a lifetime, accounting up to eight different phases of active smoking habits. Cumulative cigarette smoking was quantified in packyears, defined as the number of packs of cigarettes smoked per day multiplied by the number of years the individual has smoked.

For the present study we analyzed polymorphisms in the genes *NAT1*, *NAT2*, *CYP1A1*, *CYP1B1*, as well as the *GSTM1* and *GSTT1* deletion polymorphisms. Specifically, *NAT1* and *CYP1B1* genotypes were coded as the number of *NAT1*\*10 and *CYP1B1*\*3 alleles, respectively. *NAT2* was coded rapid acetylating conditional on the presence of at least one *NAT2*\*4 allele, which is characterized by the absence of four point mutations (as previously described in [3]), and slow acetylating otherwise. *CYP1A1* was either homozygote for the wild-type allele *CYP1A1*\*1, defined by the absence of three point mutations (rs1056827, rs1056836, rs1800440), or otherwise. *GSTM1* and *GSTT1* were characterized by the absence of their gene product.



Analyses were adjusted for age, family history of breast cancer in terms of number of affected first-degree relatives, and menopausal status classified as either pre- or postmenopausal, or unknown for women with previous hysterectomy not accompanied by bilateral oophorectomy. Menopausal status was assigned according to the reported state a year before the reference date. Study region showed no effect in an earlier analysis [5] and was hence not considered in the model.

Non-missing genetic and epidemiologic data were available for a total of 654 cases and 1085 controls. A description of the study population and of the genetic variables is given in Table 1.

## Statistical Methods

### General remarks

Interactions were only considered if all constituting main effects were in the model. We further restricted

the domain of possible gene-gene interactions to polymorphisms in the same pathway. All variables were treated as continuous and were centered on 0.

### Logistic regression and backward selection

All logistic regression models contained terms for age, family history, and menopausal status. We tested (i) main effects of smoking and the six polymorphisms separately, (ii) interaction of smoking with each polymorphism, and (iii) gene-gene interactions for polymorphisms in the same pathway. Stepwise regression with backwards selection was used to identify subsets of variables that best explained the data according to the Akaike Information Criterion [6]. We applied stepwise regression to (i) the model containing smoking and all six polymorphisms, and (ii) the model containing all main effects as well as interactions of smoking with all polymorphisms and all gene-gene interactions within pathways.

**Table 1 Study characteristics of the breast cancer case-control study in Germany and variable definition.**

Variable	Definition	Levels	Cases 654	Controls 1085
Age			42.5 +/- 5.7	42.6 +/- 5.7
Family history	first-degree relatives with breast cancer	none at least 1	87.80% 12.20%	94.80% 5.20%
Menopausal status <sup>a</sup>		premenopausal postmenopausal unknown	78.70% 6.30% 15.00%	80.60% 6.50% 12.90%
Smoking	packyears over a lifetime		8.27 +/- 12.20	6.96 +/- 10.88
NAT1	number *10 alleles	0 1 2	64.10% 32.30% 3.70%	68.60% 28.60% 2.90%
CYP1B1	number *3 alleles	0 1 2	31.00% 51.70% 17.30%	28.90% 50.90% 20.20%
NAT2	presence of at least one *4 allele	fast acetylator slow acetylator	43.70% 56.30%	39.80% 60.20%
CYP1A1	homozygote for *1 allele	Yes No	75.80% 24.20%	75.10% 24.90%
GSTT1	absence of gene product	No Yes	84.30% 15.70%	81.90% 18.10%
GSTM1	absence of gene product	No Yes	45.70% 54.30%	48.80% 51.20%

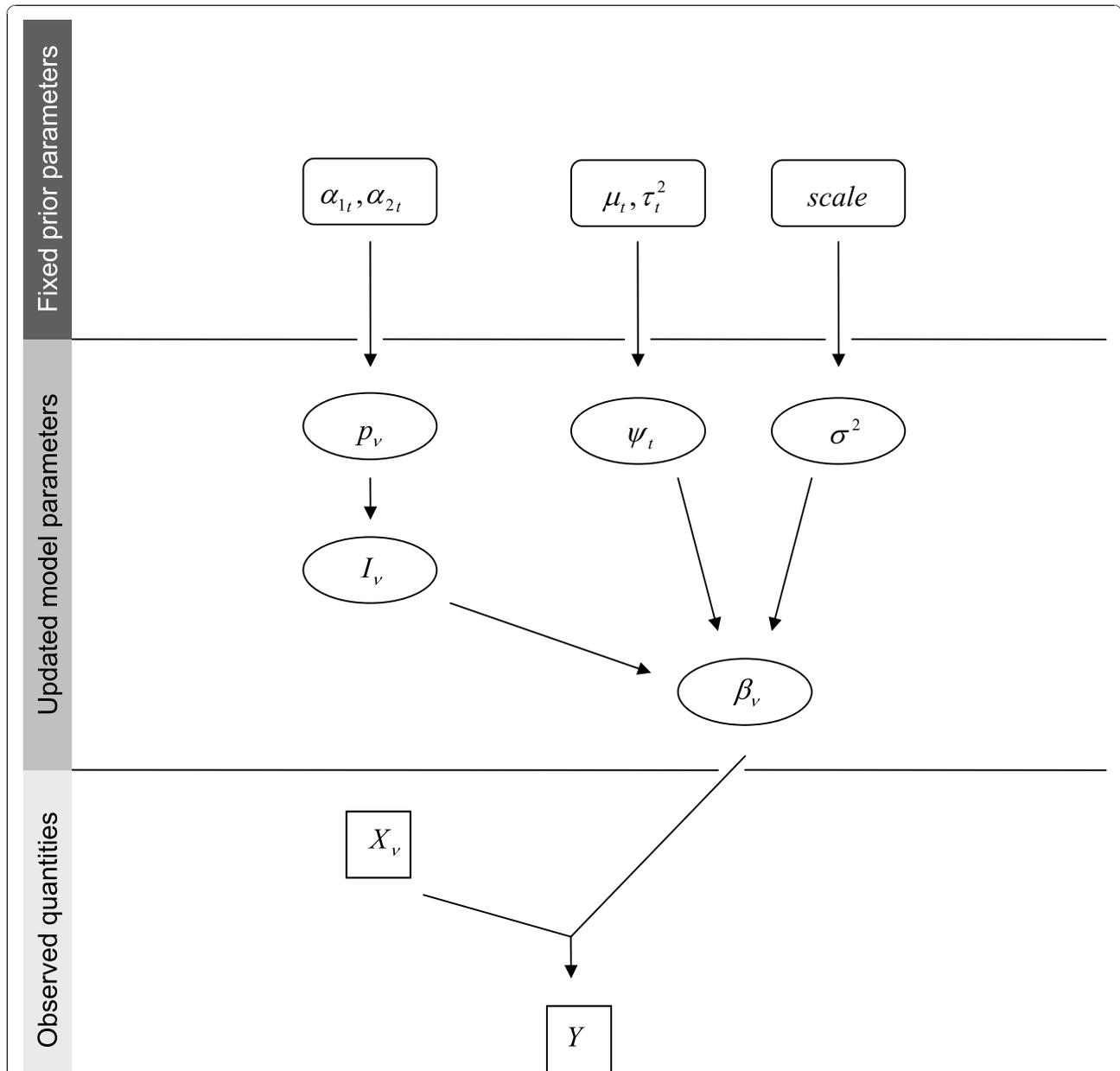
<sup>a</sup> Women with a hysterectomy not accompanied by bilateral oophorectomy were classified as unknown

**Bayes Model Averaging (BMA)**

The Bayesian model we used is illustrated as a directed acyclic graph in Figure 2. At each iteration step, we considered a logistic model of the form

$$\text{logit}(\Pr(Y = 1) | X) = \sum_c \beta_c X_c + \sum_v I_v \beta_v X_v,$$

where  $\{X_c\}$  consisted of the terms for age, family history and menopause that were included in each model.  $\{X_v\}$  contained the terms for smoking, the six polymorphisms, as well as interactions of smoking with all polymorphisms, and all gene-gene interactions within pathways. Following Conti et al. [1],  $I_v$  was a binary indicator marking the presence of term  $X_v$  in the model.



**Figure 2 Directed Acyclic Graph for BMA and its parameters.** Boxes represent observed quantities, ovals parameters to be updated over the course of MCMC, and rounded boxes fixed meta-parameters.  $Y$  denotes the dependent variable, and  $v$  indexes the sets  $X$  of independent predictor variables and  $\beta$  of corresponding estimates.  $I$  indicates inclusion of the  $v$ th variable and is Bernoulli-distributed with parameter  $p_v$ , which, in turn, follows a beta distribution with parameters  $(a_t, b_t)$  depending on the interaction level  $t$  of the variable. The variance of the coefficients  $\beta_v$  is modeled by a residual variance term  $\sigma^2$  following a half-Cauchy prior, and a variance inflation factor  $\psi_t$  depending on the interaction level and following a log-normal distribution with mean  $\mu_t$  and variance  $\tau_t$ .

Assuming that  $X_v = X_{rs}$  was an interaction term with constituting main effects  $X_r$  and  $X_s$ , then  $I_{rs} = 0$  if any of  $I_r$  or  $I_s$  were 0, formalizing the requirement that all main effects had to be in the model for an interaction to be present.

The following prior distributions were specified for the model parameters. The probability  $p_v = \Pr(I_v = 1)$  was beta-distributed with parameters  $(\alpha_{1_{main}}, \alpha_{2_{main}}) = (1, 3)$  if  $X_v$  was a main effect. This prior corresponded to a marginal inclusion probability of 0.25 for main effects and was chosen to reflect our prior emphasis on models with fewer main effects. Moreover, since the inclusion of interactions was limited by the hierarchical dependency on the presence of the main effects, we encouraged inclusion of interaction terms by specifying a greater marginal probability of including term  $X_v = X_{rs}$ , provided that  $I_r = I_s = 1$ . Specifically, we set  $(\alpha_{1_{int}}, \alpha_{2_{int}}) = (2, 2)$  and therefore, the conditional prior probability of including term  $X_{rs}$  was  $E(I_{rs} = 1 | I_r = I_s = 1) = 0.5$ , and the marginal prior probability was  $E(I_{rs} = 1) = 0.03 = 0.25 \times 0.25 \times 0.5$ .

A prior for the coefficients  $\beta_v$  was specified via

$$\beta_v \sim \begin{cases} N(0, (\psi_t \sigma)^2) & I_v = 1 \\ 0 & I_v = 0 \end{cases}$$

Thus, if  $I_v = 1$ , the variance of  $\beta_v$  was the product of a fixed component  $\sigma^2$  and of a component  $\psi_t^2$ ,  $t \in \{main, int\}$  depending on whether  $X_v$  was a main effect or an interaction. Specifically, we set  $\psi_{main}^2 = 1$ , such that  $\sigma^2$  corresponded to the variance in main effect coefficients, and  $\psi_{int}^2$  modeled the change in that variance for interaction terms. To allow for updating via the Gibbs sampler when no interaction term was present in the current model, a slightly informative prior was chosen for  $\psi_{int}^2$  via  $\log(\psi_{int}) \sim N(2.3, 1)$ .

An uninformative prior distribution was placed on  $\sigma^2$ . Since an inverse gamma prior, often chosen for the sake of its conditional conjugacy, leads to improper posterior distributions and sensitivity of inference to hyperparameter choice [7], we specified a half-Cauchy prior distribution with scale 100, as proposed by Gelman [7].

Note that via the above prior choices we a priori distinguished main and interaction effects via their model inclusion probabilities and the assumed variance of their coefficients. However, within these two groups variables were treated as exchangeable due to lack of prior evidence suggesting otherwise.

The approach was implemented using the software WinBUGS [8], running 20,000 iterations and discarding

the first 2,000 as burn-in to ensure independence of the results from the initial values. Another 2,000 iterations were discarded through the default settings of WinBUGS to allow the Markov Chain to converge; hence inference was based on 16,000 iterations. We visually inspected plots of the sampled values to ensure convergence of the chain (data not shown). WinBUGS code for our analysis is provided [see Additional file 1].

## Results

### Logistic regression

Significant associations in univariate logistic regression analysis were found for packyears and *NAT1*, as indicated by  $OR_{packyears} = 1.08$ , ( $p = 0.04$ , 95%-CI = 1.00-1.16) and  $OR_{NAT1} = 1.21$ , ( $p = 0.04$ , 95%-CI = 1.01-1.44). We also tested interactions together with their main effects and found evidence for the interaction of packyears with *NAT2* slow acetylator status ( $OR_{packyears \times NAT2} = 1.19$ ,  $p = 0.02$ , 95%-CI = 1.03-1.38). Estimates did not change substantially when the model contained all main effects (Model<sub>main</sub> in Table 2), or when all pairwise interactions (Model<sub>all</sub>) were included. Regression with backwards selection retained packyears, *NAT1*, *GSTM1* and *CYP1B1* as explanatory variables when applied to Model<sub>main</sub>, and in addition the interactions of packyears with *NAT2* and *GSTM1* when applied to Model<sub>all</sub> (results not shown). For clarity, Table 2 only tabulates interactions that showed positive findings from any of the analysis approaches.

### BMA

For BMA, posterior odds ratios and 95%-confidence intervals (CI) were calculated based on the coefficients  $\beta_v$  across all models (marginal odds ratios) and across all models with  $I_v = 1$  (conditional odds ratios). We also report the posterior probability that a variable is included in the model, expecting that variables harboring an association with disease will be included more frequently. Significance of findings was assessed via Bayes factors (BF), the ratio of posterior to prior odds that a variable was included in the model [9]. Thus evaluation of support of a non-zero coefficient took into account the specified prior distribution. Two model Bayes factors were computed in a similar fashion to evaluate support of a selected model versus (i) competing models, and (ii) the null model.

In our prior specifications we emphasized sparse models, resulting in coefficient estimates of 0 for many terms. Hence, expected values of marginal posterior odds ratios showed considerable shrinkage towards 1 with tight confidence intervals due to the hierarchical model (Table 2). The largest effect was observed for *NAT1* with  $OR_{NAT1} = 1.05$  (95%-CI = 1.05-1.05). To

**Table 2 Selected results from logistic regression and BMA.**

Variable	Logistic regression				BMA		
	Pointwise OR (95% CI)	Model <sub>main</sub> <sup>a</sup> OR (95% CI)	Model <sub>all</sub> <sup>b</sup> OR (95% CI)	OR (95% CI) <sup>&amp;</sup>	OR <sub>I = 1</sub> (95% CI) <sup>§</sup>	Pr <sub>post</sub>	BF(I = 1)
<b>packyears</b>	<b>1.08 (1.00-1.16)*</b>	<b>1.08 (1.01-1.16)*</b>	<b>1.09 (1.01-1.17)*</b>	<b>1.01 (1.01-1.01)</b>	<b>1.08 (1.08-1.08)</b>	<b>0.13</b>	<b>0.43</b>
<b>NAT1 *10</b>	<b>1.21 (1.01-1.44)*</b>	<b>1.18 (0.98-1.42)<sup>#</sup></b>	<b>1.21 (1.00-1.46)<sup>#</sup></b>	<b>1.05 (1.05-1.05)</b>	<b>1.19 (1.19-1.20)</b>	<b>0.26</b>	<b>1.05</b>
NAT2 slow vs fast <sup>1</sup>	0.86 (0.70-1.04)	0.90 (0.73-1.11)	0.89 (0.72-1.09)	0.98 (0.98-0.98)	0.87 (0.87-0.88)	0.13	0.45
CYP1A1 *1 <sup>2</sup>	0.96 (0.77-1.21)	0.94 (0.74-1.18)	0.92 (0.73-1.16)	1.00 (1.00-1.00)	0.96 (0.96-0.97)	0.06	0.19
GSTT1 deletion	0.84 (0.65-1.10)	0.86 (0.66-1.12)	0.86 (0.66-1.13)	0.98 (0.98-0.98)	0.86 (0.86-0.86)	0.12	0.42
GSTM1 deletion	1.13 (0.93-1.37)	1.15 (0.94-1.40)	1.16 (0.95-1.42)	1.01 (1.01-1.01)	1.12 (1.12-1.13)	0.09	0.31
CYP1B1 *3	0.90 (0.78-1.03)	0.90 (0.78-1.04)	0.90 (0.78-1.04)	0.99 (0.99-0.99)	0.91 (0.90-0.91)	0.09	0.31
<b>packyears × NAT2</b>	<b>1.19 (1.03-1.38)*</b>		<b>1.20 (1.03-1.40)*</b>	<b>1.00 (1.00-1.00)</b>	<b>1.19 (1.18-1.20)</b>	<b>0.01</b>	<b>0.30</b>
packyears × GSTM1	1.11 (0.96-1.28)		1.11 (0.96-1.29)	1.00 (1.00-1.00)	1.10 (1.08-1.12)	0.00	0.07

Adjusted for age, menopause and family history

<sup>1</sup> slow acetylators had at least one \*4 allele

<sup>2</sup> the reference was homozygote for the \*1 allele

<sup>a</sup> only main effects

<sup>b</sup> all main effects and interactions

<sup>&</sup> OR and CI computed from mean coefficient estimate and its standard error averaged over all models

<sup>§</sup> OR and CI computed from mean coefficient estimate and its standard error averaged over all models containing the respective variable

<sup>#</sup> p < 0.1, \* p < 0.05

assess magnitude of effects conditional on model inclusion we also tabulate expected odds ratios and confidence intervals conditional on model inclusion. The resulting values were similar to the three logistic regression scenarios, but again with much tighter confidence intervals.

The latent indicator variable *I* was used to compute the posterior probability of model inclusion for each variable. The most frequently selected predictors were *NAT1* (Pr<sub>posterior</sub> = 0.26), *NAT2* (Pr<sub>posterior</sub> = 0.13) and packyears (Pr<sub>posterior</sub> = 0.13). The posterior probability for the interaction term of packyears and *NAT2* was decreased at 0.01, due to the additional restriction that both main effects had to be present in the model.

We used Bayes factors (BFs), the ratio of posterior and prior odds that a variable was selected into the model, to assess the significance of a result in relation to the prior that had been assumed before the analysis. The following calibration has been proposed by Kass and Raftery [9] to interpret Bayes factors: between 1 and 3 suggests very mild evidence, between 3 and 20 positive evidence, between 20 and 150 strong, and above 150 very strong evidence for an association. Based on these guidelines, very mild evidence was found for *NAT1* (BF = 1.05), while all other terms exhibited Bayes factors below 1.

On the model level we first computed the Bayes factor BF<sub>all</sub> for a specific model against all remaining models to assess whether that model was superior to the competing models. Secondly, a Bayes factor BF<sub>0</sub> was computed comparing the model to the NULL model, measuring

whether any additional insight was gained in relation to the model that included only age, family history and menopausal status. To facilitate interpretation, the prior and posterior odds used in the calculation of the Bayes factors are reported along with the Bayes factors in Table 3.

In the comparison of one model to all remaining ones, we found positive evidence for the null model (BF<sub>all</sub> = 3.8) and for the model containing only *NAT1* (BF<sub>all</sub> = 3.4). Very mild evidence was suggested for the single-effect models of packyears (BF<sub>all</sub> = 1.4), *NAT2* (BF<sub>all</sub> = 1.3) and *GSTT1* (BF<sub>all</sub> = 1.1). Very mild evidence was

**Table 3 Model results for selected models.**

Model <i>M</i>	Posterior <sup>1</sup>	Prior <sup>2</sup>	BF <sub>all</sub> <sup>3</sup>	BF <sub>0</sub> <sup>4</sup>
<b>NULL</b>	<b>0.59</b>	<b>0.154</b>	<b>3.8</b>	<b>1.0</b>
packyears	0.067	0.047	1.4	0.1
<b>NAT1</b>	<b>0.159</b>	<b>0.047</b>	<b>3.4</b>	<b>0.3</b>
NAT2	0.063	0.047	1.3	0.1
CYP1A1	0.024	0.047	0.5	0.0
GSTT1	0.054	0.047	1.1	0.1
GSTM1	0.038	0.047	0.8	0.1
CYP1B1	0.044	0.047	0.9	0.1
NAT1, GSTT1	0.021	0.015	1.4	0.0
NAT1, GSTM1	0.017	0.015	1.1	0.0
NAT1, CYP1B1	0.017	0.015	1.1	0.0

Adjusted for age, menopause and family history

<sup>1</sup> obtained from MCMC-sampling

<sup>2</sup> uses BMA parameters

<sup>3</sup> support for *M* against all other models

<sup>4</sup> support for *M* against null model

also found for the combination of *NAT1* with each *GSTT1*, *GSTM1*, and *CYP1B1* ( $BF_{all} = 1.4, 1.1, \text{ and } 1.1$ , respectively). When taking into account interactions, none of the models exhibited a Bayes factor greater than 1. The same was true for Bayes factors versus the null model.

**Sensitivity analysis**

We investigated the sensitivity of our results to the choice of priors by considering different values for the prior hyperparameters described above. Specifically, we varied the values of  $(\alpha_{1_{main}}, \alpha_{2_{main}})$  and  $(\alpha_{1_{int}}, \alpha_{2_{int}})$  in the prior of  $p_v = \Pr(I_v = 1)$  to consider different expected prior probabilities of model inclusion. Moreover, different specification of  $\mu_{int}$  in the prior of  $\psi_{int}$ , as well as of the scale of the half-Cauchy prior for  $\sigma^2$ , were evaluated. Table 4 shows the different hyperparameter choices.

Estimates of posterior odds ratios showed little variation for different expected prior values of  $p_v$ . The posterior probability of model inclusion changed according to the changes in the prior parameters, i.e. doubling the prior probability of including an effect typically led to twice the posterior probability of actually including it.

Varying the mean  $\mu_{int}$  in the log-normal distribution of  $\psi_{int}$  showed no effect on the results, neither did choosing a different scale of the prior for  $\sigma^2$ .

**Discussion**

Both logistic regression and BMA highlighted a significant effect of *NAT1*. Furthermore, logistic regression showed significant effects of packyears and of the interaction of packyears with *NAT2* on breast cancer risk. The role of *NAT1* as strongest effect is supported by the Bayesian analysis of selected models. Stepwise regression analysis indicated the additional involvement of *CYP1B1* and of the interaction of packyears and *GSTM1* in breast carcinogenesis.

On a biological level, *NAT1* was initially implicated in breast cancer susceptibility through a report of a positive association of the *NAT1\*11* allele with breast cancer

risk as well as combined effects with cigarette smoking and meat consumption [10], which was, however, not confirmed in a subsequent study [11]. The inconsistent results could be attributed to sample size requirements necessary for assessing effects of *NAT1\*11*, which occurs in approximately only 3% of the general population [12]. We studied the *NAT1\*10* allele, which occurs with much greater frequency in the Caucasian population than the *NAT1\*11* allele, and may be rapid acetylating. *NAT1\*10* has been reported to be associated with higher NAT1 activity in both bladder and colon tissue [13-15]. However, the association between the *NAT1\*10* allele and increased NAT1 activity *in vivo* has not been confirmed in other studies [16-18]. For breast cancer, no significant effect of *NAT1\*10* has been found in several studies [10,11,19].

Detection of a gene effect with odds ratio in the order of magnitude that we have found for *NAT1* with 80% power at a significance level of 0.05 (assuming allele frequency 0.17, population risk 10%, log-additive disease model and unmatched 1:2 case-control design) requires 1,088 cases and twice the number of controls [20]. Thus the previous studies, as well as our own, would not have enough power to consistently detect such an effect.

Our results from logistic regression analysis regarding the association of *NAT2* with breast cancer risk, as previously reported [4], are in line with findings from other studies. In a meta- and pooled analysis including 13 studies, *NAT2* was not independently associated with breast cancer risk but smoking was found to be associated with increased risk in *NAT2* slow acetylators but not in rapid acetylators [21].

The *GSTM1* null genotype has not been found to confer susceptibility to breast cancer [22]. However, smokers carrying the *GSTM1* null genotype were at significantly elevated risk for breast cancer overall in a meta-analysis of seven studies [23]. An earlier pooled analysis of another seven smaller almost non-overlapping studies, however, did not show clear effect modification in the association between *GSTM1* and smoking [22]. Our results from stepwise regression showed a non-significant effect modification by *GSTM1*, with higher risk of breast cancer associated with smoking among those with the *GSTM1* null genotype.

Results from regression and Bayesian analyses differed in that univariate BMA analysis identified only *NAT1* as significant and did not yield significant findings for packyears and the interaction of packyears and *NAT2*. One possible explanation is that inference from BMA is based on posterior and prior probabilities instead of p-values. Thereby it avoids the problem of multiple comparisons inherent in pointwise testing of coefficients in a logistic model. In fact, none of the findings from logistic regression remain significant when Bonferroni-corrected

**Table 4 Hyperparameter scenarios for the sensitivity analysis.**

Variation of $p_v$ :			
$E(p_{main})$	0.10	0.25	0.50
$E(p_{int})$	0.25	0.50	0.75
Variation of $\psi_{int}$ :			
$\mu_{int}$	2.3	3.0	4.6
Variation of $\sigma^2$ :			
scale( $\sigma^2$ )	25	100	

for multiple testing. However, there is no simple one-to-one correspondence between frequentist and Bayesian analyses, since the latter explicitly depend on the specified priors. Our results were stable for different hyperparameter choices. However, adequate prior specification always needs to be kept in mind before starting any Bayesian analysis. In our case, mostly uninformative prior distributions were specified to reflect the lack of sufficient external information justifying an a priori distinction of variables. We therefore allowed the data more weight versus prior information in estimation and model selection. However, if desired, BMA provides a framework for the explicit inclusion of biological prior information, like pathway characteristics, into the analysis through prior specification. If one is confident about biological prior information, stronger prior assumptions may be helpful to guide the analysis. However, bias will be introduced at the same time, so that this trade-off must be carefully considered.

## Conclusions

The strength of BMA is its explicit statement of the prior assumptions given by the prior distributions for model parameters, and its consideration of model uncertainty by obtaining results averaged over a multitude of possible models. It evaluates single variables and a range of models at the same time, yielding stabilized estimates based on a set of potential data-generating models. Moreover, it provides a means to reduce dimensionality and avoids the problem of multiple comparisons.

In our study both BMA and regression analyses yielded a significant effect of *NAT1\*10*, while BMA attenuated other significant findings from logistic regression. Since all Bayesian inference depends on the specified prior information, prior choice must be carefully considered when conducting a Bayesian analysis.

## Additional material

**Additional file 1: WinBUGS code for the Bayes Model Averaging analysis.** The WinBUGS code for the Bayes Model Averaging analysis of our data, along with all used parameters.

## Acknowledgements

NS is funded by the DKFZ International PhD Program and LB by DFG grant BE 3906/2-2. The authors thank two anonymous reviewers for helpful suggestions on the manuscript.

## Authors' contributions

NS performed the statistical analyses and drafted the manuscript. LB participated in the statistical analyses and helped draft the manuscript. JCC conceived of the study, and participated in its design and coordination and helped draft the manuscript. All authors read and approved the final manuscript.

## Competing interests

The authors declare that they have no competing interests.

Received: 26 May 2010 Accepted: 16 November 2010

Published: 16 November 2010

## References

1. Conti DV, Cortessis V, Molitor J, Thomas DC: **Bayesian modeling of complex metabolic pathways.** *Hum Hered* 2003, **56**:83-93.
2. Greenland S: **Bayesian perspectives for epidemiological research. II. Regression analysis.** *Int J Epidemiol* 2007, **36**:195-202.
3. Chang-Claude J, Kropp S, Jager B, Bartsch H, Risch A: **Differential effect of NAT2 on the association between active and passive smoke exposure and breast cancer risk.** *Cancer Epidemiol Biomarkers Prev* 2002, **11**:698-704.
4. Kropp S: **The Association between Tobacco Exposure and Breast Cancer Risk considering potentially modifying Effects of specific Genotypes.** *PhD Thesis* University of Heidelberg; 2001.
5. Kropp S, Chang-Claude J: **Active and passive smoking and risk of breast cancer by age 50 years among German women.** *Am J Epidemiol* 2002, **156**:616-626.
6. Hastie TJ, Pregibon D: **Generalized linear models.** 1992.
7. Gelman A: **Prior distributions for variance parameters in hierarchical models.** *Bayesian Analysis* 2006, **1**:515-533.
8. Spiegelhalter DJ: **WinBUGS version 2.10.** 2005.
9. Raftery AE: **Bayes Factors.** *J Am Statist Assoc* 1995, **90**:773-795.
10. Zheng W, Deitz AC, Campbell DR, Wen WQ, Cerhan JR, Sellers TA, Folsom AR, Hein DW: **N-acetyltransferase 1 genetic polymorphism, cigarette smoking, well-done meat intake, and breast cancer risk.** *Cancer Epidemiol Biomarkers Prev* 1999, **8**:233-239.
11. Millikan RC: **NAT1\*10 and NAT1\*11 polymorphisms and breast cancer risk.** *Cancer Epidemiol Biomarkers Prev* 2000, **9**:217-219.
12. Loktionov A, Moore W, Spencer SP, Vorster H, Nell T, O'Neill IK, Bingham SA, Cummings JH: **Differences in N-acetylation genotypes between Caucasians and Black South Africans: implications for cancer prevention.** *Cancer Detect Prev* 2002, **26**:15-22.
13. Hughes NC, Janezic SA, McQueen KL, Jewett MA, Castranio T, Bell DA, Grant DM: **Identification and characterization of variant alleles of human acetyltransferase NAT1 with defective function using p-aminosalicylate as an in-vivo and in-vitro probe.** *Pharmacogenetics* 1998, **8**:55-66.
14. Payton MA, Sim E: **Genotyping human arylamine N-acetyltransferase type 1 (NAT1): the identification of two novel allelic variants.** *Biochem Pharmacol* 1998, **55**:361-366.
15. Bruhn C, Brockmoller J, Cascorbi I, Roots I, Borchert HH: **Correlation between genotype and phenotype of the human arylamine N-acetyltransferase type 1 (NAT1).** *Biochem Pharmacol* 1999, **58**:1759-1764.
16. Yang M, Katoh T, Delongchamp R, Ozawa S, Kohshi K, Kawamoto T: **Relationship between NAT1 genotype and phenotype in a Japanese population.** *Pharmacogenetics* 2000, **10**:225-232.
17. Badawi AF, Hirvonen A, Bell DA, Lang NP, Kadlubar FF: **Role of aromatic amine acetyltransferases, NAT1 and NAT2, in carcinogen-DNA adduct formation in the human urinary bladder.** *Cancer Res* 1995, **55**:5230-5237.
18. Bell DA, Stephens EA, Castranio T, Umbach DM, Watson M, Deakin M, Elder J, Hendrickse C, Duncan H, Strange RC: **Polyadenylation polymorphism in the acetyltransferase 1 gene (NAT1) increases risk of colorectal cancer.** *Cancer Res* 1995, **55**:3537-3542.
19. van der Hel OL, Bueno-de-Mesquita HB, van Gils CH, Roest M, Slothouber B, Grobbee DE, Peeters PH: **Cumulative genetic defects in carcinogen metabolism may increase breast cancer risk (The Netherlands).** *Cancer Causes Control* 2005, **16**:675-681.
20. Gauderman J, Morrison J, QUANTO: **A computer program for power and sample size calculations for genetic-epidemiology studies. (1.2.4).** 2006.
21. Ambrosone CB, Kropp S, Yang J, Yao S, Shields PG, Chang-Claude J: **Cigarette smoking, N-acetyltransferase 2 genotypes, and breast cancer risk: pooled analysis and meta-analysis.** *Cancer Epidemiol Biomarkers Prev* 2008, **17**:15-26.
22. Vogl FD, Taioli E, Maugard C, Zheng W, Pinto LF, Ambrosone C, Parl FF, Nedelcheva-Kristensen V, Rebbeck TR, Brennan P, et al: **Glutathione S-transferases M1, T1, and P1 and breast cancer: a pooled analysis.** *Cancer Epidemiol Biomarkers Prev* 2004, **13**:1473-1479.

23. Terry PD, Goodman M: Is the association between cigarette smoking and breast cancer modified by genotype? A review of epidemiologic studies and meta-analysis. *Cancer Epidemiol Biomarkers Prev* 2006, **15**:602-611.

doi:10.1186/1742-5573-7-10

**Cite this article as:** Stephenson *et al.*: Carcinogen metabolism, cigarette smoking, and breast cancer risk: a Bayes model averaging approach. *Epidemiologic Perspectives & Innovations* 2010 **7**:10.

**Submit your next manuscript to BioMed Central  
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

