

Supplemental Material

Examining the joint effect of multiple risk factors using exposure risk profiles: lung cancer in non smokers

Michail Papathomas^{1,2}, John Molitor¹, Sylvia Richardson¹, Elio Riboli¹ and Paolo Vineis¹

¹ Department of Epidemiology and Biostatistics, Imperial College London, U.K.

² Department of Mathematics, Coventry University, U.K.

S1. Profile regression analysis

The model (See also Molitor et al. 2010)

Our approach consists of an *assignment sub-model*, which assigns individual profiles to clusters, and a *disease sub-model*, which links clusters of profiles to an outcome of interest via a regression model. Denote, for individual i , a covariate profile as, $\mathbf{x}_i = (x_1, \dots, x_p)$; for instance, in the Gen-Air study we may have a subject with the following profile: living on a main road, exposed to more than 50mg/m³ of PM10, exposed to more than 40mg/m³ of NO₂, low physical activity at work, etc. As in Molitor et al. (2010), profiles are clustered into groups, in accordance with its covariate profile and disease status $y_i \in \{0,1\}$, and an allocation variable,

$z_i = c$ indicates the c^{th} cluster to which individual, i , belongs. Mathematically, our

basic mixture model assignment is, $\Pr(x_i) = \sum_{c=1}^C \psi_c \prod_{p=1}^P \phi_c^p(x_{ip})$. This standard mixture model (denoted as a Dirichlet process mixture, Dahl 2006) deals with a situation where the profile belongs to one of several different clusters, $c = 1, \dots, C$ each with membership probability ψ_c . The c^{th} cluster is assigned a parameter θ_c^* so that, given θ_c^* and that $z_i = c$, the phenotype y_i is a Bernoulli variable with probability θ_c^* . Denote with $\phi_c^p(x)$ the probability that risk factor x_p takes the value x , when the individual belongs to cluster c . Given that $z_i = c$, we have that $x_{i,p}$ has a multinomial distribution with cluster specific parameters ϕ_c^p so that, $x_{i,p} \sim \text{Multinom}(1, \phi_c^p)$, where, $\phi_c^p = (\phi_c^p(1), \dots, \phi_c^p(M_p))$. Here, M_p denotes the number of categories of x_p . We assume that, a priori, $\phi_c^p \sim \text{Dirichlet}(1, \dots, 1)$, which is a flat, non-informative prior. We adopt a flexible ‘stick-breaking’ prior (Ishwaran and James 2001) on the allocation weights $\psi = (\psi_1, \dots, \psi_C)$, with a random parameter α . This prior density allows for the number of clusters to be random. We place a non-informative uniform prior on α in the $(0,10)$ interval. Finally, we set that, a priori, $\theta_c^* \sim \text{Beta}(1,1)$, a usual conjugate choice of prior distribution for Bernoulli or binomial observations. Note that we use

the letter c to denote clusters, because we refer to the grouping of the observations during the estimation process, where the clustering can vary. Below, in the ‘Inference and output’ subsection, we refer to the best representative clustering of the observations and use the letter k , as in the main manuscript.

Allowing for ordinal risk factors

We extend the approach of (Molitor et al. 2010) and allow for ordinal covariates $x_{i,p}$ by introducing ordered threshold parameters $\gamma_c^p = (\gamma_c^p(1), \dots, \gamma_c^p(M_p))$, as in (Chib and Albert 1993). Now,

$$\gamma_c^p(1) < \gamma_c^p(2) < \dots < \gamma_c^p(M_p) = \infty,$$

and,

$$P(x_{ip} = 1 | z_i = c) = P(t_c^p \leq \gamma_c^p(1)),$$

⋮

$$P(x_{ip} \leq k | z_i = c) = P(t_c^p \leq \gamma_c^p(k)),$$

⋮

$$P(x_{ip} \leq M_p | z_i = c) = P(t_c^p \leq \gamma_c^p(M_p)) = 1,$$

where $1 < k < M_p$, and $t_c^p \sim N(0,1)$. Specifying the ϕ_c^p parameter vector is equivalent

to specifying the vector of ordered thresholds γ_c^p . We assign a flat prior on γ_c^p .

Implementation of this approach within a Markov chain Monte Carlo sampling framework is not a trivial matter. It is required to sample efficiently from truncated normal distributions, and convergence can be slow, depending on the sample size. This is why we had to adopt the Metropolis-Hastings sampling framework suggested in Cowles (1996).

Inference and output

The parameters of the model are estimated using Markov chain Monte Carlo (MCMC) methods (Gilks et al. 1996). A Gibbs sampling approach is used. The cluster parameters are estimated and then the subjects are allocated to clusters conditionally on these parameters. To simplify the sampling procedure, the cluster model is approximated by setting a pre-defined maximum number of clusters C . For the Genair data set we have chosen a maximum number of 15 clusters. This was satisfactory since, throughout our analysis, the subjects typically clustered into less than five groups; see the results section in the main manuscript.

Clustering procedures suffer from the so called ‘label-switching’ problem (Richardson and Green 1997). In two different iterations, what is effectively the same cluster may have a different label, so that, what was labelled as, say, cluster one at iteration one, may be labelled as cluster three at iteration two. To observe how individuals typically cluster, in a manner invariant to ‘label-switching’, we construct, at every iteration, a score matrix so that element (i,j) is one if individuals i and j are in the same cluster, and zero otherwise. When the sampling is finished, an association matrix S is built as the average of all score matrices. So, element $S_{i,j}$ is an estimate of the probability that individuals (i,j) belong to the same cluster. This is indeed a quantity that is invariant to how each cluster is labelled. We summarise this matrix and obtain a representative

average partition z_{best} using the Partitioning Around Medoids (PAM, Kaufman and Rousseeuw 1994) approach. PAM is a robust deterministic clustering algorithm that provides an optimal partition, in the sense that it maximises a clustering score related to the distance of the observations from objects that represent the structure of the data.

Model averaging approach

Assume that z_{best} consists of K subgroups. A model averaging approach is adopted to evaluate the uncertainty associated with the characteristics of these groups. This involves running through the MCMC output, obtaining, at each iteration, an average value for the model parameters across all subjects in a certain group. For example, if the first subgroup of z_{best} contains subjects $i=\{3,4,5\}$, then, from the sampled risk effects of the first iteration, we obtain $(\theta_{z_3}^*, \theta_{z_4}^*, \theta_{z_5}^*)$. A summary $\theta_1 = f(\theta_{z_3}^*, \theta_{z_4}^*, \theta_{z_5}^*)$ is then calculated, usually the mean or the median. This is repeated for all iterations of the MCMC sampler, generating a distribution of average parameter values associated with the risk of the first subgroup in z_{best} . Similar calculations are then performed for the rest of the sub-populations in z_{best} . In the same fashion, empirical distributions for $\phi_k^p(x)$, $k = 1, \dots, K$, are also obtained. The $\phi_k^p(x)$ parameters and associated distributions describe the profile of the subjects in the K subgroups of z_{best} . If the algorithm generally puts individuals in the same cluster, these empirical distributions will tend to be relatively narrow. Conversely, if the algorithm usually puts individuals in disparate clusters, then the best clustering is less typical, and the empirical distributions will tend to be relatively wide. This approach provides the interpretability of single clustering approaches like K-means, yet exploits the rich output of the MCMC sampler to assess uncertainty for the parameters and corresponding to subgroups of the best clustering. For more details on the model averaging approach through post-processing see (Molitor et al. 2010), section 3.2

S2. Checking model fit

To compare the different approaches with respect to model fit, we always use the reduced data set with 545 subjects. We first use all 545 subjects to fit the model, and then predict the probability of disease p_i for each of the subjects. Because the proportion of cases in our sample is very small (around 10%), rather than classifying the subjects as cases or controls, we use logistic regression type residuals as a measure of how well the model fits the data. The quantity we use for our comparisons is,

$$FIT = (545)^{-1} \sum_{i=1}^{545} |y_i - p_i|$$

	CART	Profile regression	MDR	Logistic regression
FIT	0.197	0.202	0.204	0.204

The CART approach has the best performance. This is to be expected, since CART is able to generate extra nodes to accommodate specific subjects with distinct profiles and phenotypes. The other methods have similar performance.

Additional References

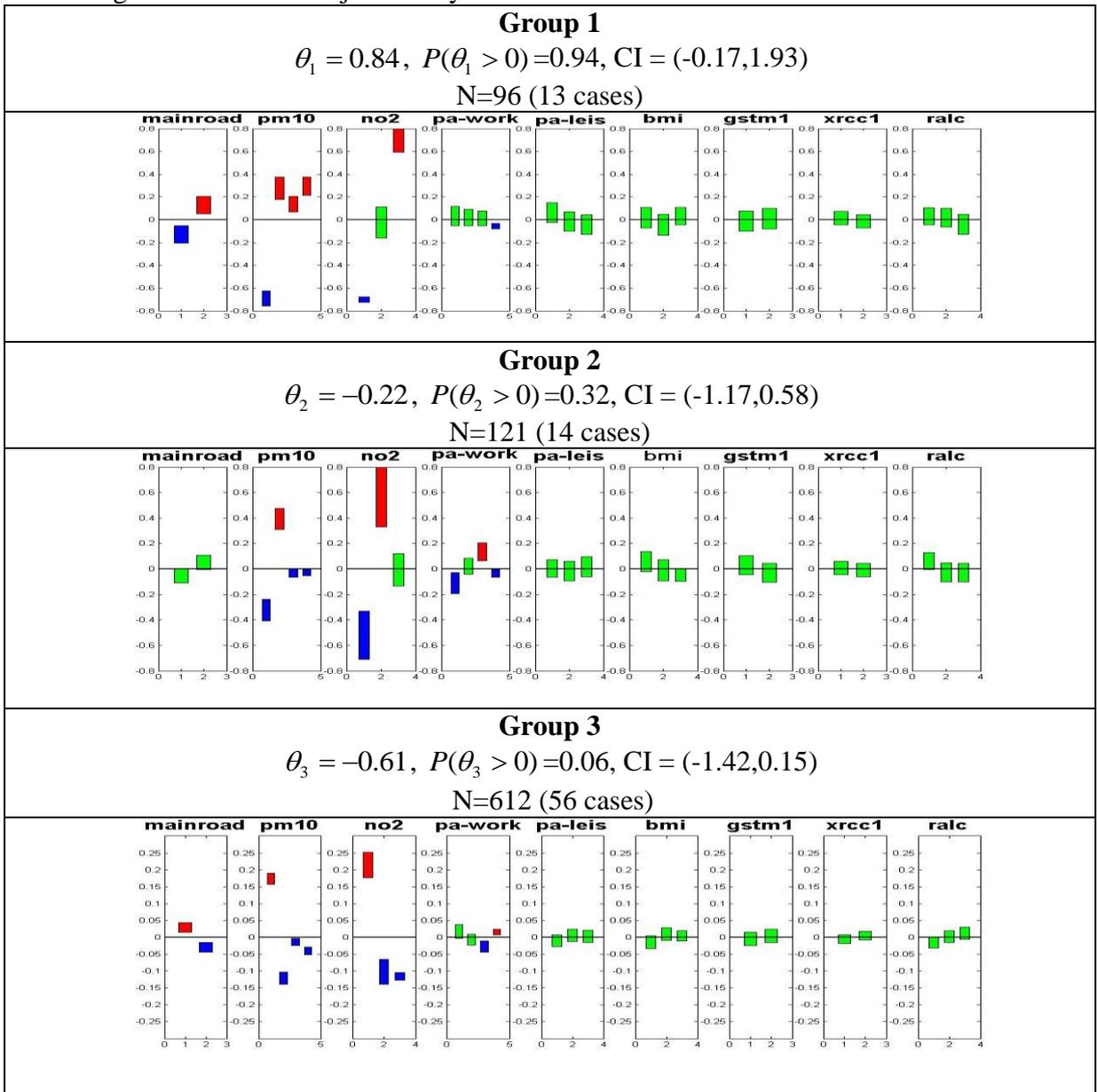
Cox TF, Cox MAA. 1994. Multidimensional Scaling. Chapman and Hall

Dahl D. 2006. Model-based clustering for expression data via a Dirichlet Process mixture model. In Bayesian Inference for Gene Expression and Proteomics. Cambridge University Press; 201-218.

Kaufman L, Rousseeuw PJ. 2005. Finding groups in data: an introduction to cluster analysis. Wiley-Interscience: Hoboken NJ.

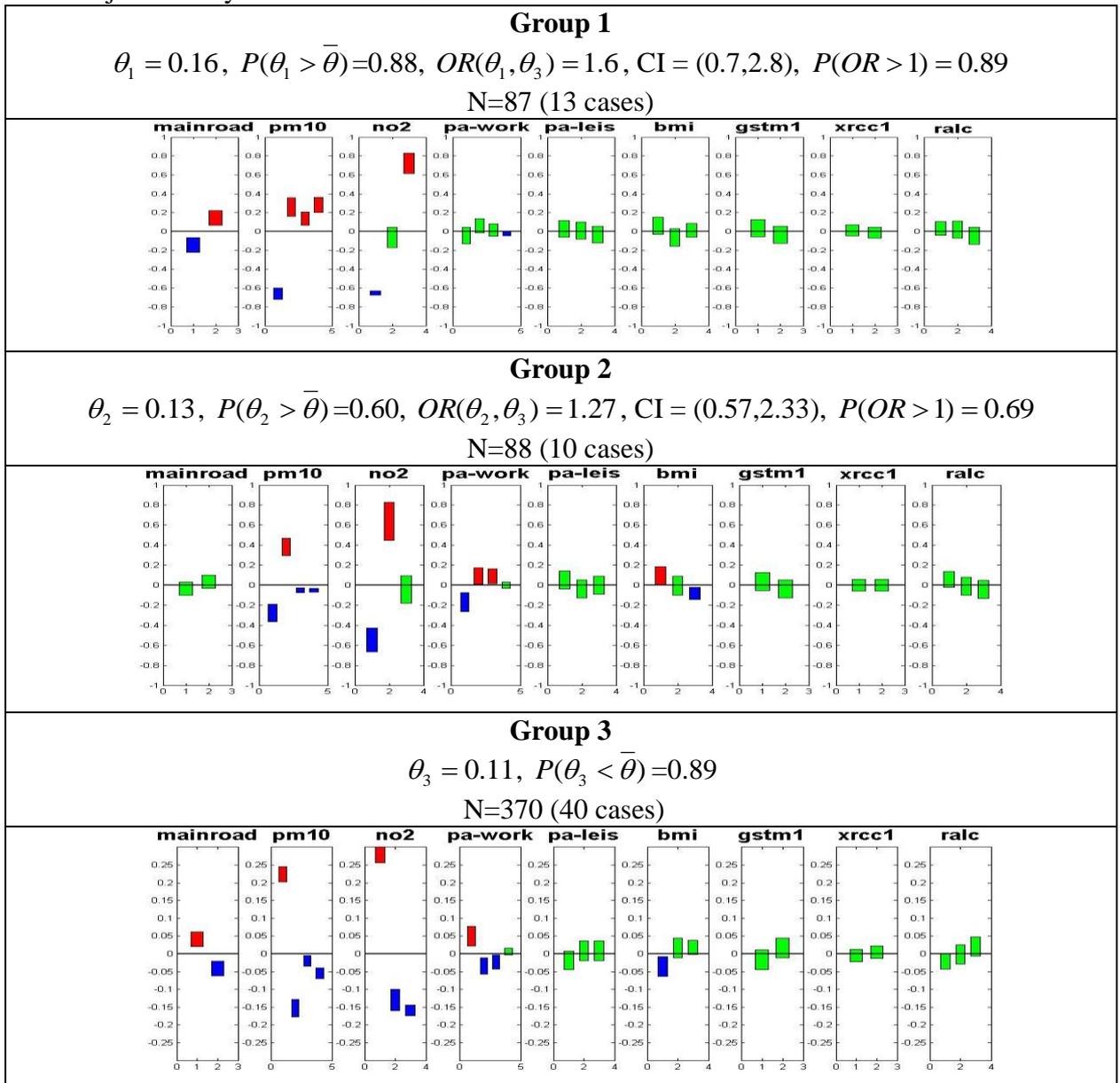
Richardson S, Green PJ. 1997. On Bayesian analysis of mixtures with an unknown number of components. J Roy Statist Soc B 59:31-792.

Supplemental Material, Table 1: Profile regression output when adjusting for the matching variables. 829 subjects analysed.



Supplemental Material, Table 2: Profile regression output with reduced data set.

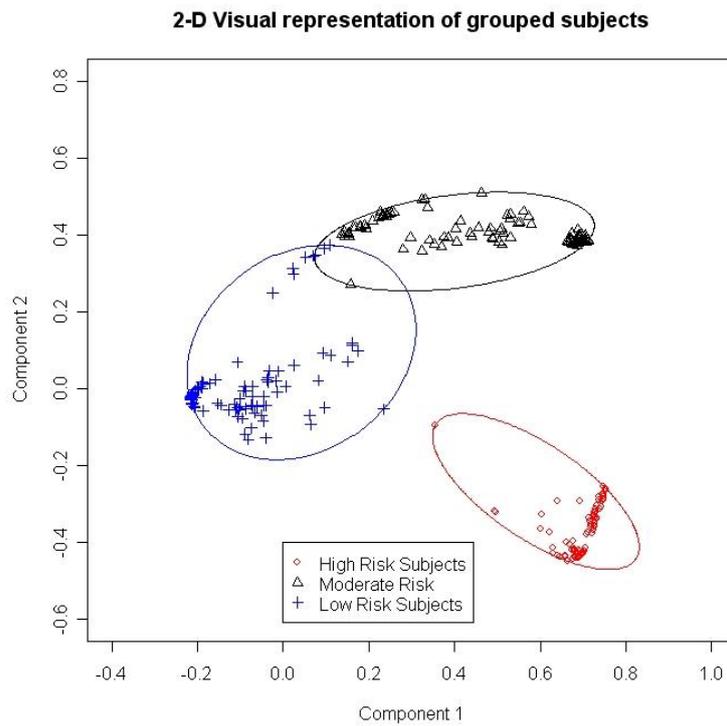
545 subjects¹ analysed with $\bar{\theta} = 0.115$



¹Due to missing PM10 observations, a reduced data set is analysed with methods other than profile regression. This reduced data set has 545 subjects (63 lung cancer cases and 482 controls) rather than 829 subjects. When analysing this data set with profile regression, results are almost identical to the results presented in Table 2 in the main manuscript. The association of the subgroups with the risk for lung cancer is less pronounced than in the full data set, but in exactly the same direction.

Supplemental Material, Table 3: p-values when each risk factor is introduced separately in a simple logistic regression analysis with an intercept. 829 subjects analysed. We present the estimated odds ratio ($p_{\text{effect}}/(1-p_{\text{effect}})/(p_{\text{ref level}}(1-p_{\text{ref level}}))$) for a specified reference level, and the associated 95% confidence interval.

	p-value	Reference level	Effect parameter	Estimated odds ratio
Living on a main road	0.36	Not on main road	On a main road	1.33 (0.72,2.45)
Exposure to PM10	0.18	<30 µg/m ³	30-40	1.10 (0.54,2.23)
			>40-50	2.27 (0.87,5.90)
			>50 µg/m ³	2.17 (0.84,5.64)
Exposure to NO ₂	0.26	<30 µg/m ³	30 – 40	1.23 (0.67,2.46)
			>40 µg/m ³	1.67 (0.88,3.15)
Physical activity at work	0.09	Sedentary occupation	Standing occupation	0.46 (0.23,0.92)
			Manual work	0.61 (0.29,1.27)
			Heavy manual work	1.15 (0.52,2.55)
Physical activity at leisure	0.14	Low	Medium	1.11 (0.59,2.07)
			High	1.66 (0.94,2.95)
Body mass index	0.45	Normal weight	Overweight	0.98 (0.60,1.60)
			Obese	0.64 (0.31,1.32)
Deletion polymorphism in <i>GSTM1</i>	0.67	Wild type	Deletion polymorphism	0.90 (0.57,1.43)
Polymorphism in <i>XRCC1</i>	0.91	Wild type	Heterozygous or homozygous variant	0.96 (0.48,1.93)
Bulky DNA adducts	0.18	Not detectable	Below median	1.99 (0.94,4.25)
			Above median	1.55 (0.73,3.31)



Supplemental Material, Figure 1: A visual 2-dimensional representation of the clustering presented in Table 2 in main manuscript. (30.51% of the distances variability is explained) This representation is constructed using the Partitioning Around Medoids (Kaufman and Rousseeuw 1994) clustering information and multi-dimensional scaling (Cox and Cox 1994).