Analysis of the effect of occupational exposure to asbestos based on threshold regression modeling of case–control data

ANTOINE CHAMBAZ*
Modal’X, Université Paris Ouest Nanterre, 92001 Nanterre
achambaz@u-paris10.fr

DOMINIQUE CHOUDAT
Département de médecine du travail, Assistance Publique-Hôpitaux de Paris, Université Paris Descartes, Sorbonne Paris Cité, 75014 Paris, France

CATHERINE HUBER
MAP5, Université Paris Descartes, Sorbonne Paris Cité, 75006 Paris, France

JEAN-CLAUDE PAIRON
Service de pneumologie et pathologie professionnelle, Centre Hospitalier Intercommunal, Faculté de médecine, INSERM U955, Université Paris-Est, 94000 Créteil, France

MARK J. VAN DER LAAN
Department of Statistics, Division of Biostatistics, School of Public Health, University of California, Berkeley, CA 94720-7360, USA

SUMMARY

We analyze the effect of occupational exposure to asbestos on the occurrence of lung cancer based on a recent French case–control (CC) study. We build a large collection of threshold regression models, data-adaptively select a better model by CC-weighted likelihood-based cross-validation and then fit this better model by CC-weighted maximum likelihood. The CC-weighting allows us to draw valid inferences from CC data without relying on a logistic regression. This is possible because the joint distribution of the indicator of being a case and matching variable is available beforehand owing to two studies independent from our data set. The implications of the fitted model in terms of years of life free of lung cancer lost due to the exposure to asbestos are discussed.

Keywords: Case–control (CC) study; Cross-validation; Occupational exposure to asbestos; Threshold regression model (TR).

*To whom correspondence should be addressed.
Asbestos is a powerful carcinogen (IARC, 1977). We rely on a recent French matched case–control (CC) study on lung cancer by Pairon and others (2009) to investigate the effect of occupational exposure to asbestos on the occurrence of lung cancer. Following a CC study design is convenient for a disease like lung cancer, with a known prevalence proportion approximately equal to five cases out of 10,000 persons (Belot and others, 2008). We model the time to incidence of lung cancer as the time until an unobservable process crosses a threshold. The effect of occupational exposure to asbestos is included by accelerating the time index of the process. The effect of other covariates is not included as a time acceleration factor, but directly on the other process parameters (mean and drift). By separately modeling the effect of covariates and exposure in this way, we can get new insights on the effect of these covariates and also, more importantly, on the effect of exposure alone. We express the latter in terms of years of life (free of lung cancer) lost (due to occupational exposure to asbestos). Our model belongs to the family of threshold regression (TR) models, which has been playing an important role in survival analysis for some years (Lee and Whitmore, 2006, 2010, and references therein). Because our study involves a description of the exposure to asbestos into 28 categories, we actually consider our model as a maximal model containing thousands of smaller models. We manage to data-adaptively select a better model in our large collection of TR models by relying on multi-fold CC-weighted likelihood-based cross-validation (van der Vaart and others, 2006). Then we fit the latter better model to the data by CC-weighted maximum likelihood (ML), and draw our conclusions from its description. The CC-weighting allows us to draw valid inferences from CC data without restricting ourselves to the logistic regression. This is made possible because the joint distribution of the indicator of being a case and matching variable is known beforehand owing to two studies independent from our data set (van der Laan, 2008).

The article is organized as follows. The data set is described in Section 2. We develop the TR modeling in Section 3. We discuss its causal interpretation (identifiability, transportability) in Section 4. In Section 5, we elaborate our inference methodology, whose application yields the results summarized in Section 6. A brief discussion is finally developed in Section 7. Some relevant material is gathered in supplementary material available at Biostatistics online.

2. **Data set**

2.1 *A matched CC study*

The matched CC study took place between 1999 and 2002 in four Parisian hospitals. Case and control subjects were retrospectively recruited at the end of each year 1999–2002 among the patients of these hospitals who were free of lung cancer at the beginning of the corresponding year. The case subjects were diagnosed with incident lung cancer during the period of the study. We assume that we can neglect truncation, i.e. the possibility that some incident cases could not participate in the study because incident cases were not immediately interviewed on diagnosis. Cases were matched by control subjects on the basis of gender, age at end of calendar year (up to ±2.5 years), hospital, and race. Control subjects were recruited among patients of the departments of ophthalmology, and general and orthopedic surgery, and were by definition free of lung cancer at the time of their enrollment. We exclude every subject with missing information. The resulting data set counts $n = 860$ cases and 901 controls, hence a total of $n + 901 = 1761$ observations. The matching pattern and race are not available. We propose an artificial valid matching pattern (based on gender, age, and hospital) and make sure that our results do not depend on this particular choice.

The population from which observations are sampled is assumed stationary during the whole study. Therefore, the pairs of case and matched controls can be modeled as independent and identically distributed random variables.
Table 1. Overall number of employments (nb. emp.) associated to each possible p/f/i description (ε), and overall number of employments featuring a particular value of the p/f/i description

<table>
<thead>
<tr>
<th>ε</th>
<th>nb. emp.</th>
<th>ε</th>
<th>nb. emp.</th>
<th>ε</th>
<th>nb. emp.</th>
<th>ε</th>
<th>nb. emp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>213</td>
<td>211</td>
<td>53</td>
<td>311</td>
<td>138</td>
<td>1..</td>
<td>582</td>
</tr>
<tr>
<td>112</td>
<td>167</td>
<td>212</td>
<td>64</td>
<td>312</td>
<td>105</td>
<td>2..</td>
<td>223</td>
</tr>
<tr>
<td>113</td>
<td>3</td>
<td>213</td>
<td>6</td>
<td>313</td>
<td>24</td>
<td>3..</td>
<td>618</td>
</tr>
<tr>
<td>121</td>
<td>150</td>
<td>221</td>
<td>59</td>
<td>321</td>
<td>136</td>
<td>.1.</td>
<td>773</td>
</tr>
<tr>
<td>122</td>
<td>46</td>
<td>222</td>
<td>36</td>
<td>322</td>
<td>189</td>
<td>.2.</td>
<td>644</td>
</tr>
<tr>
<td>123</td>
<td>3</td>
<td>223</td>
<td>3</td>
<td>323</td>
<td>22</td>
<td>.3.</td>
<td>6</td>
</tr>
<tr>
<td>131</td>
<td>0</td>
<td>231</td>
<td>2</td>
<td>331</td>
<td>1</td>
<td>..1</td>
<td>752</td>
</tr>
<tr>
<td>132</td>
<td>0</td>
<td>232</td>
<td>0</td>
<td>332</td>
<td>3</td>
<td>..2</td>
<td>610</td>
</tr>
<tr>
<td>133</td>
<td>0</td>
<td>233</td>
<td>0</td>
<td>333</td>
<td>0</td>
<td>..3</td>
<td>61</td>
</tr>
</tbody>
</table>

The total number of employments is 8432. Only 1423 of them feature a description in \( \mathcal{E} \setminus \{0\} \).

2.2 Non-occupational information

For every participant in the study, \( X \) is the age at interview and \( T \) is the age at incident lung cancer diagnosis (\( T = \infty \) if no lung cancer ever occurs). For convenience, we define \( Z = \min\{T, X\} \) and \( Y = \mathbb{1}(T \leq X) \), the indicator of being a case (\( Y = 1 \), \( Z = T \)) rather than a control (\( Y = 0 \), \( Z = X \)). The following non-occupational information is also available: \( W_0 \), the hospital of recruitment; \( W_1 \), the gender (0 for men and 1 for women); \( W_2 \), a binary indicator of occurrence of lung cancer in close family (0 if no lung cancer occurred and 1 otherwise); \( W_3 \), the discretized lifetime tobacco use (\( W_3 = 0 \) for never-smoker, \( W_3 = 1 \) for <25 pack years, \( W_3 = 2 \) for 25–45 pack years, \( W_3 = 3 \) for >45 pack years). The boundaries are chosen to yield strata of comparable sizes (371 subjects with \( W_3 = 0 \), and, respectively, 468, 469, 453 subjects with \( W_3 = 1, 2, 3 \)). We ignore the duration of habit and years of abstinence, although they are known to play a role in the development of lung cancer (Ruano-Ravina and others, 2003). We define \( W = (W_1, W_2, W_3) \in \mathcal{W} \) the explanatory covariate. The categorical matching variable \( V \in \mathcal{V} \) regroups \((W_0, W_1)\) and a discretized version of \( X \) over bins of length 5 years.

2.3 Occupational exposure to asbestos

The occupational history up to age \( X \), denoted by \( \tilde{A}(X) \), was also determined during the interview, using a “probability/frequency/intensity” (“p/f/i” for short) description characterized by the very experts who later conducted the interviews (Pairon and others, 2009). Every employment with duration at least 6 months is associated with its start and end dates, and with a p/f/i triplet summarizing the exposure to asbestos. Each coordinate of the triplet takes its values in \( \{1, 2, 3\} \) as it is evaluated as low, mild, or high. Specifically, a probability index equal to 1, 2, or 3 corresponds to a passive exposure, a possible direct exposure, or a very likely or certain direct exposure, respectively. A frequency index equal to 1, 2, or 3 corresponds to exposures occurring less than once a month, more than once a month, and during less than half of the monthly working hours or during more than half of the monthly working hours, respectively. An intensity index equal to 1, 2, or 3 corresponds to a concentration of asbestos fibers <0.1 f/mL, between 0.1 and 1 f/mL, and >1 f/mL, respectively (f/mL stands for “asbestos fibers per milliliter”). At the time of the characterization of the p/f/i description, the French threshold limit value for exposure to asbestos was set to 0.1 f/mL. We add a category 0 for no exposure at all. Hence, the set \( \mathcal{E} \) of categories of exposure has \( 27 + 1 = 28 \) elements, each of them corresponding to a particular rate of exposure.

We report in the top table of Table 1 the overall number of employments associated to each possible p/f/i description. Although computed over a total of 8432 employments, this table contains many
zeros, showing that the latter description is over-parameterized. We also report in the bottom table of Table 1 the overall number of employments that feature a particular value of each coordinate of the $p/i/v$ description.

A longitudinal history of occupational exposure to asbestos is a function $\bar{a} : [0, \infty) \to \mathcal{E}$ whose value at $t$, denoted by $a(t)$, is such that $a(t) = 0$ for $t$ small enough (before the age at first employment) or for $t$ large enough (after the age at retirement or when no further information is available; this constraint is just a convenience, as seen in Section 3). The function $\bar{a}(t)$ is characterized by $\bar{a}(t)(s) = a(s) \mathbf{1}[s \leq t]$ for all $s \geq 0$. We define $\bar{a} = 0$ iff $a(t) = 0$ for all $t \geq 0$.

One of the central issues we deal with in this article is associating each description $\epsilon \in \mathcal{E}$ with a rate of exposure $M(\epsilon)$. We employ the expression “rate of exposure” because a product $M(\epsilon) \times t$ will be interpreted as, for instance, the cumulative exposure over a period of duration $t$ at constant exposure level $\epsilon$. We propose an original solution which heavily exploits the assumed underlying multiplicative nature of the $p/i/v$ encoding, see Section 3.2.

3. TR MODELING

3.1 Time to incident lung cancer and years of life lost

We model health (relative to lung cancer) as a process starting from an initial positive value $h$ and decreasing, as time goes by, with a negative slope $\mu$, the addition of a Brownian motion $B$ taking care of the random nature of the phenomenon. In the absence of occupational exposure to asbestos, the time to incident lung cancer is modeled as the first time health reaches the threshold 0. In the presence of occupational exposure to asbestos, we model its noxious effect as an acceleration of the time scale of the health process, and the time to incident lung cancer as the first time health reaches 0 along the accelerated time scale. We choose a TR model because the related interpretation in terms of health and different time scales appeals to a wide audience, and because many biological mechanisms (e.g. neuronal and cardiac activities, ovulation...) are known to obey a threshold phenomenon. The accelerated time scale can be interpreted as a biological time scale as opposed to the reference chronological/calendar time scale.

Formally, in the absence of occupational exposure to asbestos, health at time $t \geq 0$ and time to incident lung cancer are $\inf\{h + \mu s + B_s : s \leq t\}$ and $T[h, \mu] = \inf\{t \geq 0 : h + \mu t + B_t \leq 0\}$, respectively. The distribution of $T[h, \mu]$ is known as the inverse Gaussian distribution with parameter $(h, \mu)$ (Chhikara and Folks, 1989). It holds that $T[h, \mu] < \infty$ almost surely eventually because $\mu \leq 0$. Moreover, the mean of $T[h, \mu]$ equals $h/|\mu|$ whenever $\mu < 0$.

In Section 3.2, we show how to associate a history of occupational exposure to asbestos with a non-decreasing and differentiable acceleration function $R : [0, \infty) \to [0, \infty)$, such that $R(t) \geq t$ for all $t \geq 0$. Then health at time $t \geq 0$ and time to incident lung cancer are $\inf\{h + \mu s + B_s : s \leq R(t)\}$ and $T[h, \mu, R] = \inf\{t \geq 0 : h + \mu R(t) + B_{R(t)} \leq 0\}$, respectively. Obviously, $T[h, \mu, R] = T[h, \mu]$ when $R$ is the identity, but in general $T[h, \mu, R] \leq T[h, \mu]$. Furthermore,

$$T[h, \mu, R] \geq t \iff T[h, \mu] \geq R(t), \text{ hence } T[h, \mu] = R(T[h, \mu, R]).$$  (3.1)

In light of (3.1), the time to incident lung cancer in the absence of occupational exposure to asbestos, $T[h, \mu]$, can be deduced from the history of occupational exposure to asbestos, summarized by $R$, and age at incident lung cancer, $T[h, \mu, R]$. We thus define the difference $R(T[h, \mu, R]) - T[h, \mu, R] \geq 0$ as the number of years of life free of lung cancer lost (“years of life lost” for short) that a subject could have enjoyed had the subject not been exposed to asbestos. Note that it is not necessarily a remaining number of years of life free of lung cancer, as death may occur anytime after $T[h, \mu, R]$ even in the absence of occupational exposure to asbestos. We discuss further this notion in Section 4.2.
3.2 Modeling the ageing acceleration due to occupational exposure to asbestos

We model the ageing acceleration due to occupational exposure to asbestos in two steps. First, we define a parametric model indexed by $M$ for the sake of associating each exposure level $\varepsilon$ with a rate of exposure $M(\varepsilon) \geq 1$. Second, we show how this parametric model allows us to associate every history of occupational exposure to asbestos $\bar{\varepsilon}$ with an acceleration function $R(M, \bar{\varepsilon})$.

**Rates of exposure.** The nature of the p/f/i encoding is assumed multiplicative. So, we set the rate of exposure $R$ achieves a fraction $a$ of the rate $M(\varepsilon)$ for the sake of associating each exposure level $\varepsilon$ with an acceleration function $R(M, \bar{\varepsilon})$.

**Acceleration functions.** Consider now $M \in \mathcal{M}$ and a generic longitudinal history $\bar{\varepsilon}$ as presented in Section 2. The mapping $t \mapsto M(\varepsilon(t))$ is piecewise constant. We denote by $r(M, \bar{\varepsilon})$ a continuous approximation to it. (Formally: for every point of discontinuity $t_k$ of $t \mapsto M(\varepsilon(t))$, we replace the jump by a segment joining the two points $(t_k, M(\varepsilon(t_k^−)))$ and $(t_k + \delta, M(\varepsilon(t_k + \delta)))$, where $\delta = 1/365$, leading to $r(M, \bar{\varepsilon})$.) This gives rise to the non-decreasing and differentiable acceleration function $R(M, \bar{\varepsilon}) : [0, \infty) \rightarrow [0, \infty)$ characterized by

$$R(M, \bar{\varepsilon})(t) = \int_0^t r(M, \bar{\varepsilon})(s) \, ds \quad (t \geq 0). \quad (3.3)$$

The quantity $R(M, \bar{\varepsilon})(t)$ is a summary measure of the history up to age $t$ of occupational exposure to asbestos, interpretable as the cumulative occupational exposure to asbestos up to age $t$. In particular if $\bar{\varepsilon} = \bar{0}$ (i.e. in the absence of occupational exposure to asbestos throughout a lifetime) or if $M_0 = 0$ (i.e. assuming that exposure to asbestos has no effect on lung cancer), then $R(M, \bar{\varepsilon})(t) = t$ for all $t \geq 0$: in other words, the chronological and biological time scales coincide.

Note that $R(M, \bar{\varepsilon})(t)$ is very close to a linear combination of the times spent in each job category up to age $t$, where the coefficients of the combination depend through $M$ on the job categories. In this light, our acceleration function is classical (Lee and Whitmore, 2006). What makes it original though is how $M$ maps each description $\varepsilon \in \mathcal{E}$ to its coefficient $M(\varepsilon)$. We comment further on our acceleration function in Section 3.3.

3.3 Parameterization

As explained in Section 2.1, the sampling took place at times $\tau_0, \tau_1 = \tau_0 + 1, \tau_2 = \tau_0 + 2, \tau_3 = \tau_0 + 3$ (where $\tau_0$ stands for the initial sampling date, January 1, 2000). Had the representative sampling been carried out, we would have observed $n_0$ (respectively, $n_1, n_2, n_3$) independent copies of $O^* = (W, X, A(X), Y, Z)$ sampled at time point $\tau_0$ (respectively, $\tau_1, \tau_2, \tau_3$) under, say, $P^*(\tau_0)$ (respectively, $P^*(\tau_1), P^*(\tau_2), P^*(\tau_3)$). We make the following stationarity assumption: $P^*(\tau_k) = P^*(\tau_0) = P_0^*$ for all $1 \leq k \leq 3$. We emphasize that $P_0^*$ is the distribution over the entire population sampled, not over the CC data actually obtained. Furthermore, the stationarity assumption is justified by the influx and outflow experienced by the population of the Parisian region over the course of the 4-year study.

We now build on Sections 3.1 and 3.2 to parameterize partially the distribution $P_0^*$. We assume that the conditional distribution of $T$ given $(W, X, A(X))$ equals the distribution of $T|\{h, \mu, R_3\}$, where the health...
parameter \( h \) depends on gender and the occurrence of cancer in close family \((W_1, W_2)\), the drift parameter \( \mu \) depends on \((W_1, W_2)\) and also on lifetime tobacco use \( W_3 \), and the acceleration function \( R_\lambda \) depends on the history \( \lambda(X) \) of exposure up to age \( X \).

Formally, we assume that there exist \( \alpha \in \mathbb{R}^4, \beta \in \mathbb{R}^{16}, \) and \( M \in \mathcal{M} \) such that \( \log h = \alpha_k \) with \( k = 1 + W_1 + 2W_2 \in \{1, \ldots, 4\}, \) \( \log(-\mu) = \beta_{k'} \) with \( k' = 1 + W_1 + 2W_2 + 4W_3 \in \{1, \ldots, 16\}, \) and \( R_\lambda = R(M, \lambda(X)) \). Define \( \theta = (\alpha, \beta, M) \in \Theta = \mathbb{R}^4 \times \mathbb{R}^{16} \times \mathcal{M} \). We leave unspecified the conditional distribution of \((W, X, \lambda(X))\) given \((V, Y)\). This yields a semiparametric (SP) model \( \mathcal{P}^* = \{P_0^*: \theta \in \Theta\} \) for \( P_0^* \) because we know beforehand the joint distribution of \((V, Y)\) under \( P_0^* \) (see Section B.3 of supplementary material available at Biostatistics online for its computation based on two data sets independent from ours). We assume that there exists a common dominating measure for every \( P_0^* \) and \( P_0^* \). We denote their densities by \( p_0^* \) and \( p_0^* \), respectively. With a slight abuse of notation, \( \mathcal{P}^* \) denotes both the model and the corresponding set of densities.

We assume that the mapping \( \theta \mapsto \text{KL}(P_0^*, p_0^*) \) (where KL stands for the Kullback–Leibler divergence) achieves a unique minimum at the unique \( \theta_0 \in \text{int}(\Theta) \). In Section 5, we show how to infer \( \theta_0 \) by ML estimation based on our data set and on the prior knowledge of the distribution of \((V, Y)\) under \( P_0^* \).

Comment. Oakes (1995) carried out the study of the effect of exposure to asbestos on time until death from (not incidence of) lung cancer. Although the path that leads us to our model is quite different from his, the two models have several features in common. We discuss them in Section B.1 of supplementary material available at Biostatistics online.

4. Causal interpretation

To elaborate a causal interpretation, it is necessary to go beyond the statistical model via a causal model. Following Pearl (2000), we adopt the non-parametric (NP) structural equations model (SEM) approach, among other possible choices, and assume that the reader is familiar with it. In particular, the two models have several features in common. We discuss them in Section B.1 of supplementary material available at Biostatistics online.

4.1 Identifiability and transportability in the NP SEM

The NP SEM summarized by the causal diagram of Figure 1 encodes a set of causal assumptions on how nature produces data in the context of our study. The causal diagram notably postulates that the lifetime history of tobacco consumption plays a role only through its discretized lifetime tobacco use \( W_3 \). We could substitute the original (not discretized yet) lifetime tobacco use for \( W_3 \) without perturbing the causal diagram. On the contrary, if we had access to the longitudinal lifetime history of tobacco consumption (we do not) and wished to model its action, then the causal diagram would be dramatically changed, tobacco use being a time-dependent confounder of occupational exposure to asbestos. We think that the impact in terms or residual confounding of our (data-driven) failure to exploit the longitudinal lifetime history of tobacco consumption is stronger than the impact of our decision to substitute \( W_3 \) for the original lifetime tobacco use.

We are interested in the \( w \)-specific causal effect \( P(t|\text{do}(\tilde{a} = 0), w) \). This effect is identifiable (indeed, the second rule of do-calculus applies because \( \{W\} \) d-separates \( \tilde{A} \) and \( T \) in the causal diagram obtained by deleting all arrows emerging from \( \tilde{A} \)). Licensing transportability across two populations requires knowledge of the mechanisms by which the two populations differ. Following Pearl and Bareinboim (2012), we locate the mechanisms where structural discrepancies may take place by an arrow emitted from a background variable encapsulated in a square. Switching between the populations amounts to conditioning on different values of these variables. In particular, the absence of such an arrow pointing to \( T \) represents the assumption that the mechanism responsible for determining \( T \) based on its parents is the same in both
Fig. 1. Causal diagram. Visual summary of the NP SEM encoding a set of causal assumptions on how nature produces data in the context of our study. The NP SEM becomes an SP SEM when we parameterize some of its components (namely, $\bar{A} \rightarrow R \bar{A}$ and $R \bar{A} \rightarrow T$).

populations. We see that $P(t|do(\bar{a} = \bar{0}), w)$ is directly transportable, i.e. the $w$-specific causal effect is the same in both populations (indeed, $\{\bar{A}, W\} d$-separates the set of background variables $\{S_W, S_X, S_{\bar{A}}\}$ and $T$ in the causal diagram obtained by deleting all arrows emerging from $\bar{A}$, which is a sufficient condition (Pearl and Barcainboim, 2012, Section 4)). However, we explain in Section 5 why this good news is mitigated by difficulties arising in the inference process.

4.2 Interesting features of the SP SEM

In Section 3 we have parametrically specified some components of the causal diagram, namely the causal mechanisms $\bar{A} \rightarrow R_{\bar{A}}$ (parameter $M$), $W \rightarrow T$ (parameter $(\alpha, \beta)$), and $R_{\bar{A}} \rightarrow T$ (TR, parameter $(h, \mu)$ and $U_T = B$). As a consequence, the time to incident lung cancer under the intervention $do(\bar{a} = \bar{0})$, say $T_0$, can be deduced from the natural time to incident lung cancer $T$ and $R_{\bar{A}}$ with $T_0 = R_{\bar{A}}(T)$ (see (3.1)). This motivates the presence in Figure 1 of the node $Y_{R_{\bar{A}}(T)}$ which, contrarily to $R_{\bar{A}}(T)$, is always observed, and satisfies $Y R_{\bar{A}}(T) = R_{\bar{A}}(T)$ for a case and $Y R_{\bar{A}}(T) = 0$ for a control.

The equality $R_{\bar{A}}(T) - T = T_0 - T$ gives a causal interpretation to the years of life lost (defined in Section 3.1 as its LHS term). Furthermore, we wish to emphasize that our years of life lost sometimes coincide with the expected years of life lost as introduced by Robins and Greenland (1991) (see their Theorem 3).

5. Inference

5.1 Formalizing the CC sampling

It would have been impractical and ineffective to carry out a representative sampling scheme because the probability $P_{\bar{0}}(Y = 1)$ of being an incident case of lung cancer is very small; see Section B.3 of
supplementary material available at *Biostatistics* online. In order to recruit some cases in the sample, one would have had to sample a huge number of observations. This is the main motivation for using a CC sampling scheme, which can be described as follows: First, one samples a case by sampling \((V^1, O_1^*) = (V^1, W^1, X^1, \tilde{A}(X^1), Y^1 = 1, Z^1)\) from the conditional distribution of \((V, O^*)\) given \(Y = 1\) (the superscript “\(^1\)” refers to the fact that \(Y^1 = 1\)). Second, conditionally independently from \((V^1, O_1^*)\) given \(V^1\), one samples \(J \) controls \((V_0^j, O_0^j)^* = (V_0^j, W_0^j, X_0^j, \tilde{A}(X_0^j), Y_0^j = 0, Z_0^j)\) from the conditional distribution of \((V, O^*)\) given \(Y = 0, V_0^j = V^1\) for all \(j \leq J\) (the superscript “\(0\)” refers to the fact that \(Y_0^j = 0\) for all \(j \leq J\)). Conditional on \(V^1 = v \in V\), the ratio of the number of controls for one case, \(J/V^1\), is much smaller than the ratio \(P_0^*(Y = 0|V = v)/P_0^*(Y = 1|V = v)\) one would get in the population. This scheme results in the observed data structure \(O = ((V^1, O_1^*), (V_0^j, O_0^j)^*, j = 1, \ldots, J) \sim P_0\) whose true distribution \(P_0\) can be partially deduced from \(P_0^*\) and the two-step description above (partially only, because the description does not fully characterize the dependence structure of \((V_0^j, O_0^j)^*, j = 1, \ldots, J\)). The method naturally allows one to consider the case where \(J\) is random and thus varies per experimental unit. This permits us to exploit all our observations, even though we have less cases than controls. Note that each control is only taken into account once.

### 5.2 CC-weighted log-likelihood loss function

We refer the reader to *Lee and others (2009)* for a study of the effect of diesel exhaust exposure using a TR model and CC data. Following *van der Laan (2008)* and *Rose and van der Laan (2008)*, we rely on a CC-weighted log-likelihood loss function \(\ell\) for the density \(p_0^*\) of \(P_0^*\) under sampling of \(O \sim P_0\). For the purpose of characterizing \(\ell\), consider the following facts about the inverse Gaussian distribution. Set \(h > 0, \mu < 0\), and a non-decreasing and differentiable acceleration function \(R\). The cumulative distribution function (CDF) and density of \(T[h, \mu]\) are \(t \mapsto F(h, \mu)(t) = 1 + e^{(\mu t - h)^{-1/2}} - \Phi((\mu t + h)^{-1/2})\) and \(t \mapsto f(h, \mu)(t) = (h/(2\pi t^3)^{1/2}) \exp(-(h - |\mu t|^2)/(2t))\), respectively, where \(\Phi\) is the standard normal CDF (Chhikara and Folks, 1989). By (3.1), the CDF of \(T[h, \mu, R]\) is \(t \mapsto F(h, \mu)(R(t))\), and its density is \(t \mapsto R'(t) f(h, \mu)(R(t))\). Consequently, the conditional survival function and density of \(T[h, \mu, R]\) at \(t \geq x - 1\) given \(\{T[h, \mu, R] \geq x - 1\}\) are \(G(h, \mu, R)(t) = (1 - F(h, \mu)(R(t)))/(1 - F(h, \mu)(R(x - 1)))\) and \(g(h, \mu, R)(t) = R'(t) f(h, \mu)(R(t))/(1 - F(h, \mu)(R(x - 1)))\), respectively. The characterization of \(\ell\) also requires the definition of the following quantities, used as weights: for each \((y, v) \in [0, 1] \times V\),

\[
(q_0, q_0(y|v), q_0(v|y), \tilde{q}_0(v)) = \left( P_0^*(Y = 1), P_0^*(Y = 1|V = v), P_0^*(V = 1|Y = y), q_0 \frac{q_0(0|v)}{q_0(1|v)} \right). \quad (5.1)
\]

We are now able to proceed: for all \(p_0^* \in \mathcal{P}^*\), setting \(O_1^* = (\Omega^1, Z^1)\) and \(O_0^* = (\Omega_0^0, Z_0^{0,j})\), we define

\[
\ell(p_0^*)(O) = q_0 \log p_0^*(V^1, O_1^*) + \tilde{q}_0(V^1) \sum_{j=1}^J \log p_0^*(V_0^j, O_0^j)^* \\
= q_0 \log p_0^*(Z^1|\Omega^1, V^1) + \tilde{q}_0(V^1) \sum_{j=1}^J \log p_0^*(Z_0^j, \Omega_0^0, V^1) + \text{rem}(O) \\
= q_0 \log g(\theta)(Z^1) + \tilde{q}_0(V^1) \sum_{j=1}^J \log G(\theta)(Z_0^j) + \text{rem}(O), \quad (5.2)
\]

where \(\text{rem}(O)\) is a random term independent of \(\theta\) and with notational convention \(G(\theta)(Z) = G(h, \mu, R_\tilde{A})(Z), g(\theta)(Z) = g(h, \mu, R_\tilde{A})(Z)\) for \(\log(h) = \alpha_1 + W_1 + 2W_2, -\log(\mu) = \beta_1 + W_1 + 2W_2 + 4W_3, R_\tilde{A} = R(M, A(X))\).
We call \(\ell(p_0^*)(O)\) a “CC-weighted log-likelihood” because \(\log p_0^*(O^*)\) is the log-likelihood of \(O^*\) under \(P_0^*\), whereas \(\ell(p_0^*)(O)\) is not necessarily the log-likelihood of \(O\) under a distribution \(P_0\) deduced from \(P_0^*\) and the two-step description of Section 5.1. Even though \(q_0\) appears in both terms of the RHS expression in (5.2), we prefer to consider \(\ell(p_0^*)(O)\) as defined above rather than \(q_0^{-1}\ell(p_0^*)(O)\). This choice guarantees that \(\ell(p_0^*)(O)\) is on the same scale as \(\log p_0^*(O^*)\). The weighted log-likelihood loss function is adapted to the CC sampling scheme in the following sense (the proof is relegated to Section C of supplementary material available at Biostatistics online).

**Proposition 1** Let us assume that model \(P^*\) for \(p_0^*\) is such that \(\int \log p_0^*(o^*)dP_0^*(o^*, Y = y)\) are properly defined for all \(p_0^* \in P^*\) and \(y = 0, 1\). If \(P^*\) is well specified (i.e. if \(P_0^* \in P^*\)), then the density that maximizes the expectation under \(P_0\) of the weighted log loss function (5.2) over \(P^*\), \(\arg \max_{P_0^* \in P^*} E_{P_0} \ell(p_0^*)(O)\), is unique and coincides with \(p_0^*\).

Thus, it is possible to draw inference about \(P_0^*\) based on data sampled from \(P_0\) by maximizing the relevant part of the CC-weighted log-likelihood of the data at \(\theta \in \Theta\), which is written as \(P_n \tilde{\ell}(\theta) = \sum_{i=1}^n [q_0 \log g(\theta)(Z_i^1) + q_0 (V_i^1)(1/J) \sum_{j=1}^J \log G(\theta)(Z_i^{0,j})]\), where \(\tilde{\ell}(\theta)(O) = \ell(p_0^*)(O) - \text{rem}(O)\) (see (5.2)) and \(P_n = \sum_{i=1}^n \delta_{O_i}\) denotes the empirical measure. Following the lines of the proof of Proposition 1, the ML estimator \(\theta_n(P_n) = \arg \max_{\theta \in \Theta} P_n \tilde{\ell}(\theta)\) estimates \(\theta_0\) (the unique minimizer of \(\theta \mapsto \text{KL}(p_0^*, p_0^*)\) over \(\text{int}(\Theta)\)). We briefly consider some asymptotic properties (including consistency and asymptotic normality) in Section A of supplementary material available at Biostatistics online.

**Comment.** We showed in Section 4.1 that the \(w\)-specific causal effect \(P(t|do(a = \tilde{0}), w)\) is directly transportable. This good news is mitigated by the need to nevertheless resort to a transportability assumption. Indeed, our inference approach fundamentally relies on the key quantities (5.1) which are used as weights to eliminate the bias due to the CC sampling. Fortunately, we can compute them owing to (Belot and others, 2008), an independent study of cancer incidence and mortality in France over the period 1980–2005, and on data made publicly available by the French National Institute of Statistics and Economic studies (see Section B.3 of supplementary material available at Biostatistics online).

However, we must assume, as far as the joint distribution of \((V, Y)\) is concerned, either that the data from (Belot and others, 2008), which are collected over the whole French population, are representative of the Parisian population of interest, or that sampling from the four Parisian hospitals that participate in the study is stochastically equivalent to sampling from the population of France.

### 6. Application

#### 6.1 Fitting the best model

There is no doubt that the model we have built so far is over-dimensional. The p/vi description with its 28 different levels is itself certainly too rich (see Table 1), or at least difficult to establish and prone to errors. We rather consider the model \(\{P_0^*: \theta \in \Theta\}\) described so far as a “maximal” model giving rise to a large collection of submodels \(\{P_k^*: \theta \in \Theta_k\}\) obtained by adding constraints on the “maximal” parameter \(\theta = (\alpha, \beta, M) \in \Theta\). The number of such submodels is large indeed: there are \((1 + 7^3) = 344\) submodels defined by adding only constraints on \(M\) (of the type \(M_0 = 0\), or \(M_0 > 0\) and, for any \(k = 1, 2, 3, 0 = M_{k,1}\) or \(M_{k,1} = M_{k,2}\) or \(M_{k,2} = 1\) or \(0 = M_{k,1} = M_{k,2}\) or \(M_{k,1} = M_{k,2} = 1\) or \((0 = M_{k,1}, M_{k,2} = 1)\), hence the total number of submodels equals \(2^2 \times 2^3 \times 344 = 11 008\). It is out of question to explore the whole collection of submodels. Instead, we propose first to define a large collection \(\{\Theta_k: k \in K\}\) of interest, and second to let the data select a better \(\Theta_k\) in the latter collection based on a multi-fold likelihood-based cross-validation criterion. van der Vaart and others (2006) show that, under mild assumptions, the multi-fold
likelihood-based cross-validation criterion will select a better model comparing favorably with the oracle model of the collection (whose definition involves the true distribution of the data). By this we mean that, for every $\delta > 0$, the likelihood risk of the better model will not exceed the sum of $(1 + \delta)$ times the likelihood risk of the oracle model with an error term of the form $c(\delta) \log \log |K|/n$. Although we cannot invoke rigorously this property here, it motivates the procedure that we describe in Sections B.4 and B.5 of supplementary material available at *Biostatistics* online.

The best model $\{P_\theta^*: \theta \in \Theta_\delta\}$ is described in Section B.5 of supplementary material available at *Biostatistics* online. Its characterization teaches us that neither the initial health nor the drift parameter depend on the indicator of occurrence of lung cancer in close family. Moreover, (i) a passive exposure (probability index equal to 1) is the same as no exposure at all and (ii) being exposed less than once a month (frequency index equal to 1) or between once a month and during less than half of the monthly working hours (frequency index equal to 2) have the same effect.

We fit the best model in terms of ML on the whole data set. Regarding the derivation of confidence intervals (CIs), we decide to rely on the bootstrap instead of a central limit theorem (such as Proposition 3 in Section A of supplementary material available at *Biostatistics* online). The particulars of the bootstrap procedure follow. We set $B = 1000$ and $p = 5\%$; then for $b$ ranging from 1 to $B$, we repeatedly resample without replacement $n(1 - p) = 817$ observed data structures, yielding the bootstrapped empirical measure $P_n^{b(1-p)}$, in order to compute and store the corresponding ML estimate $\theta_{n(1-p),\delta}^*(P_n^{b(1-p)})$ of $\theta \in \Theta_\delta$. The mean and median values of $\theta_{n,\delta}^* = \{\theta_{n(1-p),\delta}^*(P_n^{b(1-p)}): b \leq B\}$ only very slightly differ from each other. Moreover, they are close to the ML estimate $\theta_n^*(P_n)$ computed on the whole data set. The componentwise $0.025\%-16$- and $(1 - 0.025\%-16)$-quantiles of $\theta_{n,\delta}^*$ are used as lower and upper bounds of CIs, which simultaneously provide a 95% coverage by the applied Bonferroni correction. The procedure yields the values reported in Table 2, which we now comment on.

The statements we make in this paragraph are simultaneously valid with a 95% probability. Regarding initial health $h$, it is seen in particular that women are associated with a significantly larger $h$ than men. Two main features arise regarding the drift $\mu$. Firstly, for each level of lifetime tobacco use, $|\mu|$ is significantly larger for men than for women (actually, the CIs for $W_5 = 3$ slightly overlap). Combined with the already mentioned fact that women are associated with a larger $h$, this implies that, for any given history of exposure to asbestos and for every level of lifetime tobacco use, the distribution of time to incident lung cancer in women is stochastically dominated by the distribution of time to incident lung cancer in men. In other words, given a man and a woman sharing the same history of exposure to asbestos and lifetime tobacco use, given an age $t$, the man is more likely to have developed an incident lung cancer at age $t$ than the woman. Note that there is no clear consensus in the literature on whether there exist differences in lung cancer risk between men and women or not (for instance, Zang and Wynder, 1996 argue that women are more susceptible to tobacco carcinogens, but Haiman and others, 2006 show that, in certain ethnic and racial group, men and women are equally susceptible to tobacco carcinogens). Secondly, both in men and women, $|\mu|$ significantly increases with lifetime tobacco use. This implies that, both in men and women, for any given history of exposure to asbestos and for every $0 \leq w < w' \leq 3$, the distribution of time to incident lung cancer for lifetime tobacco use equal to $w$ is stochastically dominated by the distribution of time to incident lung cancer for lifetime tobacco use equal to $w'$. In other words, given two persons sharing the same gender and history of exposure to asbestos, the person with the larger lifetime tobacco use is more likely to have developed an incident lung cancer at age $t$ than the other. This is in agreement with the general scientific consensus (Biesalski and others, 1998). As for parameter $M$, we see that our estimate is very different from the value of $M$ which was arbitrarily chosen, not learned from the data, by Pairon and others (2009) (namely: $M_0 = 27$ and $(M_{1,l}, M_{2,l}, M_{3,l}) = (\frac{1}{3}, \frac{2}{3}, 1)$ for $l = 1, 2, 3$). Note that Table 2 also includes the factor of acceleration of time $(M(\varepsilon) - 1)$ and related CIs for each $\varepsilon \in \mathcal{E} \setminus \{0\}$, as derived from the other estimates.
For 601 similar cases we compute a counterpart of the subtable devoted to the years of life lost. A confidence upper bound could be derived in the study, his or her individual years of life lost. Say that we mostly care for a pointwise estimation of, and confidence lower bound on, the years of life lost. In light of Section 3.1, the results of the previous section provide us with a way of evaluating, for each exposure level, the years of life lost (and corresponding 95% confidence lower bounds) for the remaining 259 cases. We report the quartiles, mean, and extreme values of years of life lost as computed on those 259 cases. We represent in Figure 2 the empirical CDF of the years of life lost (and corresponding 95% confidence lower bounds) for the remaining 259 cases. We report in Table 3 the quartiles, mean, and extreme values of years of life lost as computed on those 259 cases.

6.2 Years of life lost

In light of Section 3.1, the results of the previous section provide us with a way of evaluating, for each case in the study, his or her individual years of life lost. Say that we mostly care for a pointwise estimation of, and confidence lower bound on, the years of life lost. A confidence upper bound could be derived similarly. We compute a counterpart of the subtable devoted to the years of life lost as computed on those 259 cases. We report the quartiles, mean, and extreme values of years of life lost as computed on those 259 cases. We represent in Figure 2 the empirical CDF of the years of life lost (and corresponding 95% confidence lower bounds) for the remaining 259 cases. We report in Table 3 the quartiles, mean, and extreme values of years of life lost as computed on those 259 cases.

Table 2. Estimated values of $\theta = (\alpha, \beta, M)$ (precision $10^{-2}$) and related CIs obtained by fitting the best model

<table>
<thead>
<tr>
<th>$W_1 = 0$</th>
<th>$W_1 = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>23.82 [23.42; 24.13]</td>
<td>25.09 [24.86; 25.40]</td>
</tr>
</tbody>
</table>

$-100\mu = -100\exp(\beta_1 + W_1 \beta_2 + 4W_1)$

<table>
<thead>
<tr>
<th>$W_1 = 0$</th>
<th>$W_1 = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.69 [0.08; 1.46]</td>
<td>0.02 [0.01; 0.03]</td>
</tr>
<tr>
<td>7.70 [6.91; 8.28]</td>
<td>6.53 [5.73; 7.68]</td>
</tr>
<tr>
<td>17.67 [17.11; 18.38]</td>
<td>14.79 [13.65; 17.77]</td>
</tr>
</tbody>
</table>

$M_1$: 1.19 [0.34; 2.00]

| $M_{1.1} = 0$ | $M_{1.2}: 0.97 [0.96; 0.99]$ | $M_{1.3} = 1$ |
| $M_{2.1} = M_{2.2}$ | $M_{2.2}: 0.93 [0.90; 0.98]$ | $M_{2.3} = 1$ |
| $M_{3.1}: 0.02 [0.00; 0.09]$ | $M_{3.2}: 0.09 [0.00; 0.27]$ | $M_{3.3} = 1$ |

Resulting values of the factor of acceleration of time ($M(\varepsilon) - 1$) (precision $10^{-3}$) and related CIs for each level of exposure $\varepsilon \in \mathcal{E} \setminus \{0\}$. Recall that $M(0) = 1$. The parameters $h$ and $\mu$ do not depend on $W_2$. A Bonferroni correction ensures that the CIs simultaneously guarantee 95% coverage.
Table 3. Quartiles, mean, and extreme values of the years of life lost and corresponding 95% confidence lower bounds (precision $10^{-3}$), as computed on those 259 cases (i.e. 30% of all cases) for whom the evaluated years of life lost is positive

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>25%</th>
<th>50%</th>
<th>Mean</th>
<th>75%</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years of life lost</td>
<td>0.026</td>
<td>0.289</td>
<td>0.769</td>
<td>2.467</td>
<td>2.408</td>
<td>36.577</td>
</tr>
<tr>
<td>95% lower bound</td>
<td>0.001</td>
<td>0.014</td>
<td>0.037</td>
<td>0.555</td>
<td>0.102</td>
<td>12.832</td>
</tr>
</tbody>
</table>

Fig. 2. Empirical distributions of years of life lost and related confidence lower bound. The rightmost curve with bullets (respectively, leftmost curve with triangles) represents the empirical CDF of the years of life lost (respectively, of the 95% confidence lower bound on that number) of those cases for whom it is positive, that is, the empirical CDF of \( \delta(T_1^i, \tilde{A}_1(T_1^i)) > 0, i \leq n \) (respectively, \( \delta^- (T_1^i, \tilde{A}_1(T_1^i)) : \delta(T_1^i, \tilde{A}_1(T_1^i)) > 0, i \leq n \)). Only 30% of the cases are concerned. The x-axis scale is logarithmic.

7. Discussion

In Section 4.1, we commented on the fact that it would have been more relevant to rely on a longitudinal lifetime history of tobacco consumption than on lifetime tobacco use, but we do not have access to such comprehensive information pertaining to tobacco consumption. However, we do have more information than the four-category discretized lifetime tobacco use, \( W_3 \), that we finally exploited. The choice to rely on \( W_3 \) was mainly motivated by the wish to simplify the presentation of our approach, and by the intention to use fewer parametric components in our model. Indeed, the whole study could be easily generalized by replacing the NP models \( \log h = \alpha_1 + W_1 + 2W_2 \) and \( \log(-\mu) = \beta_1 + W_1 + 2W_3 + 4W_4 \) (see Section 3.3) with identifiable parametric models \( \log h = m_1(W' ; \alpha) \) and \( \log(-\mu) = m_2(W' ; \beta) \), where \( W' = (W_1, W_2, W_3, W_4) \) incorporates the original (not discretized yet) lifetime tobacco use, \( W_3' \), and possibly additional covariates \( W_4' \).

As shown in van der Laan (2008), we could base our study on any parametric model for the conditional distribution of the time to incidence of lung cancer given the remaining relevant information. We choose a TR model because the related interpretation in terms of health appeals to a wide audience, and because it is biologically meaningful. We emphasize that assuming that \( B \) is a Brownian motion is only one of several
Analysis of the effect of asbestos based on TR modeling

possibilities. Likewise, seeing the effect of exposure expressed as an acceleration $R$ of the time of the process is easily comprehensible, as well as the resulting useful equality $R(T[h, \mu, R]) = T[h, \mu]$. The parameter we estimate, $\theta_0 = (\alpha_0, \beta_0, M_0)$, is defined by projecting $P_0^*$ onto a TR model. It notably yields the acceleration function $(\bar{a}, t) \mapsto R(M_0, \bar{a})(t)$. A promising alternative approach would be to estimate directly $\arg\min_{M \in M} E[(T_0 - R(M, A(T))(T))^2 | Y = 1]$ (or a regularized version of it) by relying on the theory of SP estimation.

We intend to build on the present study and go further in the aforementioned directions in future work.

SUPPLEMENTARY MATERIAL


ACKNOWLEDGMENTS

A. Chambaz was a member of the MAP5 department (UMR CNRS 8145, Université Paris Descartes) when he conducted this research. He would like to thank M.-L. Ting Lee for interesting discussions on TR models, the associate editor, and two anonymous reviewers for their excellent suggestions. Conflict of Interest: None declared.

FUNDING

This work was supported by a Fulbright Research Grant and ANSES (ES 2005-006).

REFERENCES


[Received March 30, 2012; revised September 3, 2013; accepted for publication September 4, 2013]