Threshold regression models adapted to case-control studies, and the risk of lung cancer due to occupational exposure to asbestos in France

Antoine Chambaz, Dominique Choudat, Catherine Huber, Jean-Claude Pairon, Mark van Der Laan

Threshold regression models adapted to case-control studies, and the risk of lung cancer due to occupational exposure to asbestos in France

A. Chambaz[1,*], D. Choudat[2,†], C. Huber[1], J-C. Pairon[3], M. J. van der Laan[4]

[1] *MAP5, Université Paris Descartes and CNRS*

[2] *Assistance Publique – Hôpitaux de Paris and Université Paris Descartes*

[3] *INSERM U955 and Université Paris-Est Créteil*

[4] *University of California, Berkeley*

March 17, 2011

Abstract

Asbestos has been known for many years as a powerful carcinogen. Our purpose is quantify the relationship between an occupational exposure to asbestos and an increase of the risk of lung cancer. Furthermore, we wish to tackle the very delicate question of the evaluation, in subjects suffering from a lung cancer, of how much the amount of exposure to asbestos explains the occurrence of the cancer. For this purpose, we rely on a recent French case-control study. We build a large collection of threshold regression models, data-adaptively select a better model in it by multi-fold likelihood-based cross-validation, then fit the resulting better model by maximum likelihood. A necessary preliminary step to eliminate the bias due to the case-control sampling design is made possible because the probability distribution of being a case can be computed beforehand based on an independent study. The implications of the fitted model in terms of a notion of maximum number of years of life guaranteed free of lung cancer are discussed.

Keywords: case-control study; cross-validation; threshold regression model.

1 Introduction

Asbestos has been known for many years as a powerful carcinogen [1]. Our purpose is to quantify the relationship between an occupational exposure to asbestos and an increase of the risk of lung cancer. Furthermore, we wish to tackle the very delicate question of the evaluation, in subjects suffering from a lung cancer, of how much the amount of exposure to asbestos explains the occurrence of the cancer.

For this purpose, we rely on a recent French case-control study on lung cancer [9]. For a sample of approximately 2,000 participants, a number of information is available, including

[*]This collaboration took place while Antoine Chambaz was a visiting scholar at UC Berkeley, supported in part by a Fulbright Research Grant and the CNRS. A. Chambaz would like to thank M-L. Ting Lee for interesting discussions on threshold regression models.

[†]This work was supported by a grant from the Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (ES 2005-006).

information pertaining to lifetime tobacco consumption and a longitudinal description of occupational exposure to asbestos. Each employment is associated with its duration and an original qualitative description of the exposure to asbestos into 28 categories.

We decide to model the age at incident lung cancer as the first time that a time-indexed continuous stochastic process (which should be interpreted as an amount of health relative to lung cancer, initially positive and featuring a negative trend) reaches 0. This justifies the expression *first hitting time model*, but the expression *threshold regression model* is often preferred. Such models have been playing an important role in survival analysis for some years now, and we refer the reader to [6, 7] for a bibliographical overview. The model is designed in such a way that occupational exposure to asbestos may accelerate the reference time, so that incident lung cancer may occur sooner in the presence of exposure to asbestos than it would in the absence of any such exposure. This actually yields a very large collection of threshold regression models, the largest one (*i.e.*, less constrained) containing thousands of smaller threshold regression models (obtained for instance by reducing the original 28-category description of exposure to asbestos to a description with fewer categories).

As mentioned earlier, the dataset has been obtained following a case-control study design, which is convenient for a rare disease like lung cancer (since it allows to sample known cases of lung cancer). In that sense, case-control sampling is a biased sampling method. In our example, approximately one out of two participants is a case, *i.e.* is diagnosed an incident lung cancer, a proportion which is of course much larger than in the population of interest, with a known prevalence proportion approximately equal to five cases out of 10,000 persons [2]. Knowing (actually: Estimating based on the independent study [2]) beforehand the probability distribution of being a case is of crucial importance, as it makes it possible to eliminate the bias induced by the case-control sampling design, as shown in [13, 12]. Indeed, we manage to data-adaptively select a better model in our large collection of threshold regression models by relying on multi-fold likelihood-based cross-validation [15]. Then, we fit the latter better model to the data by maximum likelihood, therefore obtaining a quantitative understanding of how an exposure to asbestos is related to an increase of the risk of lung cancer.

The evaluation of how much the amount of exposure to asbestos explains the occurrence of an incident lung cancer in a case is a recurring issue. It has important implications in public-health policy-making and might be used in the design of legal compensation schemes (as in the United States, unlike in France). In this view, a mathematical notion of *probability of causation* has been formalized and studied in [10]. The authors of [10] soon overcame the shortcomings of the latter notion which they had underlined, by showing that *expected years of life lost due to hazardous exposure* can sometimes be estimated, and how to estimate them when possible [11, 8]. In this article, we explain and take advantage of the fact that resorting to threshold regression modeling makes it very easy to come up with a notion of *maximal number of years of life guaranteed free of lung cancer* (heuristically, a number of years of life which a subject living infinitely would enjoy before developing an incident lung cancer). Once the selected model has been fitted, elementary algebra maps deterministically (conditionally on observed age at incident lung cancer, history of occupational exposure, and parameter estimates) an age at incident lung cancer and a longitudinal description of occupational exposure to asbestos into a number which can be interpreted as a maximal number of years of life guaranteed free of lung cancer.

We emphasize that, although the central issues studied in this article are *causal* by their very nature, we cautiously used for their statement two expressions (how an exposure is *related*

to an increase; how much the amount of exposure *explains* the cancer) which belong to the semantic field of *associations*. This wariness is notably motivated by the fact that smoking is also a well known risk factor of lung cancer [3], so that reaching a causal conclusion would require unraveling the intertwined effects of asbestos exposure and smoking, an impossible task with the dataset at hand. For this reason among others, the above mentioned notion of maximal number of years of life guaranteed free of lung cancer cannot be interpreted causally.

The article is organized as follows. The dataset and the original qualitative description of the exposure to asbestos into 28 categories are carefully described in Section 2. The case-control estimation problem is formalized in Section 3, and some asymptotic properties of the resulting case-control weighted maximum likelihood estimator are briefly exposed in Section 4 (elements of proofs are relegated to the appendix). We develop the threshold regression modeling in Section 5. This includes the formal definition of the maximal number of years of life guaranteed free of lung cancer. Section 6 is dedicated to the application itself. This includes the computation of the quantities required to eliminate the bias due to the case-control sampling design, the details of the model selection procedure, the description of the better model fitted to the data, and its implications regarding the maximal number of years of life guaranteed free of lung cancer. A brief discussion is finally developed in Section 7.

2 Dataset

A case-control study.

The dataset was built following a case-control sampling scheme. The 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 to 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. They 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, general and orthopedic surgeries, and were by definition free of lung cancer at the time of their enrollment.

The one-to-one matching (*i.e.*, the pattern of who is matched by whom) and race are not available. We come up with an artificial valid matching pattern (based on gender, age and hospital) and make sure that our results do not depend on this particular choice. We exclude every subject with missing information. The full data set then counts $n = 860$ cases and 901 controls, resulting in $n + 901 = 1,761$ observations.

The population sampled from during the study is arguably stationary. Therefore, the observed data structures on experimental units made of pairs of case and matched control can be modeled as independent and identically distributed (iid) random variables. This simple fact is the *cornerstone* of the study undertaken here. Following the seminal article [13], we invoke this fact in order to derive the valid likelihood function which is the backbone of the study. The fundamental reasoning is fully developed with care in Section 3.

Finally, we emphasize that the results we obtain in this article, based on this dataset, can be interpreted as results relative to France under the additional assumption that sampling from the four Parisian hospitals that participate to the study is stochastically equivalent to sampling from the population of France. This assumption is also carefully stated in Section 3.

Non-professional information.

Each subject included in the study is associated with his/her date of birth, gender, date of incident lung cancer diagnosis (for cases) or interview (for controls), and binary indicator of occurrence of lung cancer in close family.

Information pertaining to tobacco consumption is also collected. We know for each subject if he/she ever smoked. For those subjects who were once smokers, the beginning and ending dates of the smoking period are given, as well as the lifetime tobacco use.

We will however summarize this information by only considering a discretized version of the lifetime tobacco use. Our motivation is twofold. First, a relevant tobacco history would be dynamic whereas we only have cumulated information. Second, such a time dependent tobacco history would yield time dependent confounding. Furthermore, the previous argument also implies that the results we obtain in this article cannot be interpreted in causal terms. It is well known that tobacco is a serious risk factor of lung cancer [3]. Reaching a causal interpretation would require that we unravel the intricate synergies between tobacco use and occupational exposures to asbestos, a difficult task that we cannot even try to address given the data at hand.

Occupational information.

Occupational information on subjects is longitudinal. Every employment (with duration at least 6 months) is associated with its start and end dates as well as with an original description of the exposure to *asbestos*, a known carcinogen.

This description is a triplet referred to as "probability/frequency/intensity", each of them taking values in $\{1, 2, 3\}$: for the considered employment, the probability of exposure, its frequency and intensity are evaluated as low/mild/high, respectively coded by 1, 2, 3. Hence, the set $\mathcal{E}$ of categories of exposure has 27+1=28 elements (we add a category $0 = (0, 0, 0)$ for no exposure), each of them corresponding to a particular rate of exposure. Note that we will use from now on either the notation $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$ or more simply the notational shortcut $\varepsilon = \varepsilon_1\varepsilon_2\varepsilon_3$.

We report in Table 1 the overall number of employments associated to each possible "probability/frequency/intensity" description. Although computed over a total of 8,432 employments, Table 1 strikingly exhibits many zeros, showing that the latter description is over-parametrized.

We also report in Table 2 the overall number of employments that feature a particular value of each coordinate of the "probability/frequency/intensity" description. Notice that, of course, the sums over rows coincide.

The generic longitudinal description of occupational exposure to asbestos is denoted by $\bar{a}$. It belongs to $\bar{\mathcal{E}}$, the set of functions from the nonnegative real line to $\mathcal{E}$ such that $a(t) = 0$ for t small or large enough (before the age at first employment or when no further information is available; this constraint is just a convenience, as we will make clear in Section 5). It is understood that the value of $\bar{a}$ at t is denoted by $a(t)$ while $\bar{a}(t)$ stand for the restrictions of $\bar{a}$ to $[0, t]$. Thus $a(t) = a$ correspond to an occupational position held at age t and characterized by asbestos exposure a.

One of the central issues we deal with in this article is how to associate each description in $\mathcal{E}$ with a rate of exposure. We propose an original solution which heavily exploits the underlying multiplicative nature of the "probability/frequency/intensity" encoding. Indeed,

ε	nb. of emp.	ε	nb. of emp.	ε	nb. of emp.
111	213	211	53	311	138
112	167	212	64	312	105
113	3	213	6	313	24
121	150	221	59	321	136
122	46	222	36	322	189
123	3	223	3	323	22
131	0	231	2	331	1
132	0	232	0	332	3
133	0	233	0	333	0

Table 1: Overall number of employments associated to each possible "probability/frequency/intensity" description. The total number of employments is 8,432. Only 1,423 of them feature a description in $\mathcal{E} \setminus \{0\}$.

	1	2	3
probability	582	223	618
frequency	773	644	6
intensity	752	610	61

Table 2: Overall number of employments that feature a particular value of each coordinate of the "probability/frequency/intensity" description.

it is the product of "probability", "frequency" and "intensity" which is relevant in terms of rate of exposure.

3 Formulation of the case-control estimation problem

This section builds upon the seminal study [13]. Following its strategy:

(i) We derive from the description of our case-control study the characterization of the prospective sampling scheme one would have liked to follow, had the probability of being an incident case of lung cancer not been so small. This mainly amounts to defining an observed data structure $O^\star$ under prospective sampling, whose distribution $P_0^\star$ presents features of interest.

(ii) In view of the latter, we characterize the observed data structure O under matched case-control sampling and its distribution P_0. Then we show how to make inference on the features of $P_0^\star$ from data sampled under P_0.

Prospective sampling.

We first set a calendar time τ (expressed in years), and consider a generic subject sampled at time τ. We denote by $W = (W_1, W_2, W_3) \in \mathcal{W}$ his/her explanatory covariate taking values in $\mathcal{W} = \{1,2,3,4\} \times \{0,1\}^2 \times \{0,1,2,3\}$,

- W_0 indicating from which hospital the generic subject is sampled;
- $W_1 = 0$ for men and $W_1 = 1$ for women;

- $W_2 = 0$ if no lung cancer occurred in close family and $W_2 = 1$ otherwise;
- $W_3 = 0$ for never-smoker, $W_3 = 1$ for lifetime tobacco use comprised between 1 and 25 pack years, $W_3 = 2$ for lifetime tobacco use comprised between 26 and 45 pack years, $W_3 = 3$ otherwise.

Note that the boundaries that we chose for defining W_3 yield strata of comparable sizes (371 subjects with $W_3 = 0$, and respectively 468, 469, 453 subjects with $W_3 = 1, 2, 3$). Let T denote his/her age at incident lung cancer (set to infinity if no lung cancer ever occurs), and let $X = X(\tau)$ denote his/her age at time τ. They are associated with $Z = \min\{T, X\}$ and $Y = \mathbf{1}\{T \leq X\}$. Finally, the occupational information collected at time τ is encoded in $\bar{A}(X)$.

Now, as explained in Section 2, sampling occurred at times $\tau_0, \tau_1 = \tau_0 + 1, \tau_2 = \tau_0 + 2, \tau_3 = \tau_0 + 3$ (where τ_0 stands for the initial sampling date, January 1st, 2000). Obviously the reference population depends on time. Denoting by $P^\star(\tau)$ the distribution of the reference population at time τ, we make the following *stationary assumption*:

$$\forall 1 \leq k \leq 3, P^\star(\tau_k) = P^\star(\tau_0) \equiv P_0^\star. \tag{1}$$

This assumption is reasonable due to the influx and the outflow featured by the population of the Parisian region over the period of investigation. We emphasized in Section 2 that interpreting the results obtained in the present article as results relative to France requires that one be willing to assume that sampling from the four Parisian hospitals that participate to the study is stochastically equivalent to sampling from the population of France. Formally, this amounts to assuming that the distribution of the observed data structure $O^\star$ sampled from the whole French population equals $P_0^\star$.

Had the prospective sampling been undertaken, we would have observed n_0 (respectively n_1, n_2, n_3) independent observed data structures $O_i^\star$ sampled at time points τ_0 (respectively τ_1, τ_2, τ_3), therefore collecting an iid sample $(O_1^\star, \ldots, O_N^\star)$ of size $N = n_0 + n_1 + n_2 + n_3$ of the distribution $P_0^\star$ by virtue of our stationary assumption. This justifies the final definition of the observed data structure in a prospective sampling scheme:

$$O^\star = (W, X, \bar{A}(X), Y, Z) \tag{2}$$

with

- W explanatory covariate;
- X the age of the subject associated with the unit when it is sampled;
- $\bar{A}(X)$ occupational information up to age X related to asbestos;
- $Y = 1$ if and only if (iff) $T = Z \leq X$ (the subject is then called a case) and $Y = 0$ iff $T > Z = X$ (the subject is then called a control).

The likelihood of $O^\star$ under $P_0^\star$ finally writes as

$$\begin{aligned} P_0^\star(O^\star) = P_0^\star(W) P_0^\star(X|W) P_0^\star(\bar{A}(X)|X, W) \\ \times dP_0^\star(Z = T | T \geq X - 1, \bar{A}(X), X, W)^Y \\ \times P_0^\star(T > X | T \geq X - 1, \bar{A}(X), X, W)^{1-Y}, \end{aligned} \tag{3}$$

where $dP_0^\star(t | T \geq X - 1, \bar{A}(X), X, W)$ is the conditional density of T at time t given the event $[T \geq X - 1, \bar{A}(X), X, W]$.

Matched case-control sampling.

Such a prospective sampling scheme would have been impractical and ineffective because the probability $P_0^\star(Y = 1)$ of being an incident case of lung cancer is very small. In order to recruit some cases in the sample, one would have to sample a huge number of observations. This is the main motivation for using a case-control sampling scheme.

Let us now describe what is our observed data structure in this framework. We introduce the categorical matching variable $V \in \mathcal{V}$ obtained by concatenating W_0 (subject's hospital when sampled), W_1 (subject's gender) and a discretized version of the age at sampling X over bins of length five years. In the sequel, we repeatedly use the convenient (though redundant) notation (V, W).

The matched case-control sampling scheme can be described as follows:

- One first samples a case by sampling

$$(V^1, O^{1\star}) = (V^1, W^1, X^1, \bar{A}^1(X^1), Y^1 = 1, Z^1)$$

  from the conditional distribution of $(V, O^\star)$ given $Y = 1$.

- Subsequently, one samples J controls

$$(V^{0,j}, O^{0,j\star}) = (V^{0,j}, W^{0,j}, X^{0,j}, \bar{A}^{0,j}(X^{0,j}), Y^{0,j} = 0, Z^{0,j})$$

  from the conditional distribution of $(V, O^\star)$ given $Y = 0, V^{0,j} = V^1$ for all $j \leq J$.

This results in the observed data structure

$$O = ((V^1, O^{1\star}), (V^{0,j}, O^{0,j\star}), j = 1, \ldots, J) \sim P_0$$

whose true distribution P_0 can be deduced from $P_0^\star$ and the two-step description above. Interestingly, the method naturally allows to consider the case that J is random and thus varies per experimental unit. This allows to exploit all our observations, even though we have less cases than controls. Note that each control is only taken into account once.

Case-control weighting of the log**-likelihood loss function developed for prospective sampling.**

It is remarkable that the log-likelihood loss function developed for prospective sampling can be adapted to the case-control sampling scheme by appropriate weighting. This weighting relies on the *prior knowledge* of the following probabilities: for each $(y, v) \in \{0, 1\} \times \mathcal{V}$,

$$q_0 = P_0^\star(Y = 1), \tag{4}$$
$$q_0(y|v) = P_0^\star(Y = y|V = v), \tag{5}$$
$$q_0(v|y) = P_0^\star(V = v|Y = y), \tag{6}$$

or, namely, the marginal probability of being a case (4), the conditional probabilities of being a case or a control given matching variable at level v (5), and the conditional probabilities of observing level v for the matching variable given being a case or a control (6). Indeed, it is possible to compute the latter key quantities based on the independent study [2], see Section 6.1.

For this purpose, let us define for all $v \in \mathcal{V}$ the quantities

$$\bar{q}_0(v) = q_0 \frac{P_0^\star(Y=0|V=v)}{P_0^\star(Y=1|V=v)} = q_0 \frac{q_0(0|v)}{q_0(1|v)} \tag{7}$$

and introduce the following case-control weighted log-likelihood loss function for the density $p_0^\star$ of $P_0^\star$ *under sampling of* $O \sim P_0$:

$$\ell(O|p^\star) = q_0 \log p^\star(V^1, O^{1\star}) + \bar{q}_0(V^1) \frac{1}{J} \sum_{j=1}^{J} \log p^\star(V^1, O^{0,j\star}). \tag{8}$$

It is worth noting that, even though q_0 appears in both terms in (8), we prefer to consider $\ell(O|p^\star)$ as defined above rather than $q_0^{-1}\ell(O|p^\star)$. This choice guarantees that the weighted log-likelihood $\ell(O|p^\star)$ is on the same scale as the log-likelihood $\log P_0^\star(O^\star)$ under prospective sampling.

Proposition 1. *Let $p_0^\star$ be the density of the observed data structure $O^\star$ under prospective sampling. Consider a model $\mathcal{P}^\star$ for $p_0^\star$ such that the integrals $\int \log p^\star(o^\star) dP_0^\star(o^\star, Y=y)$ are properly defined for all $p^\star \in \mathcal{P}^\star$ and $y = 0, 1$. If $\mathcal{P}^\star$ is well-specified (*i.e.*, if $p_0^\star \in \mathcal{P}^\star$), then the density that maximizes the expectation under P_0 of the weighted loss function (8) over $\mathcal{P}^\star$,*

$$\arg\max_{p^\star \in \mathcal{P}^\star} E_{P_0} \ell(O|p^\star),$$

is unique and coincides with $p_0^\star$.

The proof is relegated to the appendix.

In Section 5 we propose a parametric model for the conditional distribution of $O^\star$ given $(W, X, \bar{A}, Y)$, that is the conditional distribution of Z given $(W, X, \bar{A}, Y)$. The parametric model is sound in the sense that the conditional distribution of Z given $(W, X, \bar{A}, Y)$ only depends on $(W, X, \bar{A}(X), Y)$. The latter parametric model is combined with a nonparametric model for the conditional distribution of $(W, X, \bar{A})$ given Y, both yielding a semiparametric model $\mathcal{P}^\star$ for $p_0^\star$ because we know beforehand q_0, the true probability of being a case.

Let $p_\theta^\star$ be the density of $O^\star$ under parameter θ. Assuming that the true density $p_0^\star$ is "projected" (in terms of Kullback-Leibler divergence) onto $p_{\theta_0}^\star$ for some θ_0, or in other terms that the mapping $\theta \mapsto \mathrm{KL}(p_0^\star, p_\theta^\star)$ achieves a unique minimum at the unique θ_0, we focus hereafter on the maximum likelihood estimation of θ_0. Following the lines of the proof of Proposition 1, the case-control weighted maximum likelihood estimator

$$\theta_n = \arg\max_\theta \sum_{i=1}^{n} \ell(O_i|p_\theta^\star) \tag{9}$$

does estimate θ_0. We briefly consider its asymptotic properties in the next section.

4 Asymptotic properties of the case-control weighted maximum likelihood estimator

Let us denote for convenience

$$\Omega = (W, X, \bar{A}(X), Y)$$

so that $O^\star = (\Omega, Z)$, and in the same spirit,

$$\begin{aligned}\Omega^1 &= (W^1, X^1, \bar{A}^1(X^1), Y^1 = 1)\\ \Omega^{0,j} &= (W^{0,j}, X^{0,j}, \bar{A}^{0,j}(X^{0,j}), Y^{0,j} = 0),\end{aligned}$$

hence $O^{1\star} = (\Omega^1, Z^1)$ and $O^{0,j\star} = (\Omega^{0,j}, Z^{0,j\star})$. We consider a semiparametric model such that the likelihood of $O^\star$ under $\theta \in \Theta$ writes as

$$p^\star_\theta(O^\star) = p^\star_\theta(Z|\Omega)\eta(\Omega),$$

$\eta(\Omega)$ being here the likelihood of Ω, which we assume without serious loss of generality to be bounded away from 0. Therefore the weighted log-likelihood loss function for $p^\star_0$ under sampling of $O \sim P_0$ can be decomposed as

$$\begin{aligned}\ell(O|p^\star_\theta) &= q_0 \log p^\star_\theta(V^1, O^{1\star}) + \bar{q}_0(V^1)\frac{1}{J}\sum_{j=1}^J \log p^\star_\theta(V^1, O^{0,j\star})\\ &= q_0 \log p^\star_\theta(Z^1|\Omega^1, V^1) + \bar{q}_0(V^1)\frac{1}{J}\sum_{j=1}^J \log p^\star_\theta(Z^{0,j}|\Omega^{0,j}, V^1)\\ &\quad +\mathrm{rem}(O), \qquad (10)\end{aligned}$$

where $\mathrm{rem}(O)$ is a random term independent of θ. We set $\tilde{\ell}(O|\theta) = \ell(O|p^\star_\theta) - \mathrm{rem}(O)$ and note that one can substitute $\tilde{\ell}(O_i|\theta)$ for $\ell(O_i|p^\star_\theta)$ in the definition (9) of θ_n without modifying the resulting estimator:

$$\theta_n = \arg\max_{\theta\in\Theta} \sum_{i=1}^n \tilde{\ell}(O_i|\theta),$$

therefore avoiding to consider η at all while estimating the parameter of interest.

We recall that the class $\mathcal{F} = \{\ell(\cdot|p^\star_\theta) : \theta \in \Theta\}$ is P_0-Glivenko-Cantelli if $\sup_{\theta\in\Theta} |\frac{1}{n}\sum_{i=1}^n \ell(O_i|p^\star_\theta) - E_{P_0}\ell(O|p^\star_\theta)| = o_{P_0}(1)$. The following classical consistency result holds (see Theorem 5.7 and Example 19.8 in [14]).

Proposition 2. *Assume that $E_{P_0^\star} \log p^\star_0(O)$ is well-defined and that the mapping $\theta \mapsto \mathrm{KL}(p^\star_0, p^\star_\theta)$ from Θ to the nonnegative real numbers attains its minimum uniquely at $\theta_0 \in \mathrm{int}(\Theta)$ ($p^\star_{\theta_0}$ is the Kullback-Leibler projection of $p^\star_0$ upon $\{p^\star_\theta : \theta \in \Theta\}$). If $\mathcal{F}$ is P_0-Glivenko-Cantelli then θ_n converges in probability to θ_0. This is for instance the case if Θ is a compact metric space, if $\theta \mapsto \ell(o^\star|p^\star_\theta)$ is continuous for every $o^\star$ and if $\mathcal{F}$ admits an integrable envelope function with respect to P_0.*

It is accompanied with an asymptotic normality result (inspired by classical results of asymptotic normality, see Theorem 5.23 in [14]; we omit the measurability conditions).

Proposition 3 (first part)**.** *In the context of Propositions 1 and 2, assume in addition that $\theta \mapsto \log p^\star_\theta(Z|\Omega)$ is twice differentiable at θ_0 $P_0^\star$-almost surely with first and second derivatives $\dot{\ell}^\star_{\theta_0}(O^\star)$ and $\ddot{\ell}^\star_{\theta_0}(O^\star)$ such that $\int l(o^\star)dP_0^\star(o^\star, Y = y)$ are properly defined for $l = \dot{\ell}^\star_{\theta_0}, \ddot{\ell}^\star_{\theta_0}$ and $y = 0, 1$, and introduce accordingly the weighted versions*

$$\begin{aligned}\dot{\tilde{\ell}}(O|\theta_0) &= q_0\dot{\ell}^\star_{\theta_0}(V^1, O^{1\star}) + \bar{q}_0(V^1)\frac{1}{J}\sum_{j=1}^J \dot{\ell}^\star_{\theta_0}(V^1, O^{0,j\star}),\\ \ddot{\tilde{\ell}}(O|\theta_0) &= q_0\ddot{\ell}^\star_{\theta_0}(V^1, O^{1\star}) + \bar{q}_0(V^1)\frac{1}{J}\sum_{j=1}^J \ddot{\ell}^\star_{\theta_0}(V^1, O^{0,j\star}).\end{aligned}$$

Suppose also that, for every θ_1, θ_2 in a neighborhood of θ_0 and a function $\dot{m}$ such that $E_{P_0^\star}\dot{m}(O^\star)^2 < \infty$, $P_0^\star$-almost surely

$$|\log p^\star_{\theta_1}(Z|\Omega) - \log p^\star_{\theta_2}(Z|\Omega)| \leq \dot{m}(O^\star)\|\theta_1 - \theta_2\|.$$

Furthermore, assume that $\theta \mapsto E_{P_0}\ell(O|p^\star_\theta) = E_{P_0^\star}\log p^\star_\theta(O^\star)$ (by virtue of Proposition 1) admits a second-order Taylor expansion at θ_0 with nonsingular symmetric second derivative matrix $S_{\theta_0} = E_{P_0}\ddot{\tilde{\ell}}(O|\theta_0)$. Then $S_{\theta_0} = E_{P_0^\star}\ddot{\ell}^\star_{\theta_0}(O^\star)$ and

$$\sqrt{n}(\theta_n - \theta_0) = -S^{-1}_{\theta_0}\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\tilde{\ell}}(O_i|\theta_0) + o_{P_0}(1). \tag{11}$$

In particular, the sequence $\sqrt{n}(\theta_n - \theta_0)$ is asymptotically Gaussian with mean zero and covariance matrix $\Sigma = S^{-1}_{\theta_0}E_{P_0}[\dot{\tilde{\ell}}(O|\theta_0)\dot{\tilde{\ell}}(O|\theta_0)^\top]S^{-1}_{\theta_0}$.

We emphasize that we purposely do not use the same convention to denote the first and second order derivatives at θ_0 of $\theta \mapsto \log p^\star_\theta(Z|\Omega)$ (respectively $\dot{\ell}^\star_{\theta_0}(O^\star)$ and $\ddot{\ell}^\star_{\theta_0}(O^\star)$) and the derivatives of $\theta \mapsto \tilde{\ell}(O|\theta)$ (respectively $\dot{\tilde{\ell}}(O|\theta_0)$ and $\ddot{\tilde{\ell}}(O|\theta_0)$): we intend to stress that the former are related to the prospective sampling observed data structure $O^\star$ whereas the latter are related to the case-control sampling observed data structure O.

A natural question arises: How does the asymptotic covariance matrix Σ compares with the asymptotic covariance matrix one would have got under prospective sampling? We give in the second part of Proposition 3 a very simple answer, but for a prospective sampling under a modified version of $P_0^\star$. Introduce for clarity of exposition the notation

$$\bar{q}_0(o^\star) = q_0\frac{q_0(y|v)}{q_0(1|v)} \tag{12}$$

such that $\bar{q}_0(o^\star) = q_0$ if $y = 1$ ($o^\star$ corresponds to a case) and $\bar{q}_0(o^\star) = \bar{q}_0(v)$ if $y = 0$ ($o^\star$ corresponds to a control). By setting

$$\frac{dP_1^\star}{dP_0^\star}(o^\star) = \frac{1}{2\bar{q}_0(o^\star)}, \tag{13}$$

we define a probability distribution $P_1^\star$ for the observed data structure $O^\star$ (indeed, $o^\star \mapsto \bar{q}_0(o^\star)$ is positive and $E_{P_0^\star}(2\bar{q}_0(O^\star))^{-1} = 1$). Moreover:

- under $P_1^\star$, being a case is as likely as being a control (equivalently, $P_1^\star(Y = 1) = \frac{1}{2}$);
- the marginal distribution of the matching variable V under $P_1^\star$ equals the conditional distribution of V under $P_0^\star$, conditionally on being a case (equivalently, $P_1^\star(V = v) = q_0(v|1)$ for all $v \in \mathcal{V}$);
- given (V, Y), $O^\star$ has the same distribution under $P_1^\star$ as under $P_0^\star$ (indeed, $\bar{q}_0(o^\star)$ depends on $o^\star$ through (v, y) only).

Furthermore, since obviously

$$2E_{P_1^\star}\bar{q}_0(O^\star)\log p^\star_\theta(O^\star) = E_{P_0^\star}\log p^\star_\theta(O^\star),$$

$\theta \mapsto \bar{q}_0(O^\star)\log p^\star_\theta(O^\star)$ is a proper loss function for the purpose of estimating θ_0 under prospective sampling of $O^\star \sim P_1^\star$.

Proposition 3 (second part)**.** *Define* $S'_{\theta_0} = E_{P_1^\star}\bar{q}_0(O^\star)\ddot{\ell}^\star_{\theta_0}(O^\star)$*. Suppose that the model is well-specified (or equivalently that* $\mathrm{KL}(p_0^\star, p_{\theta_0}^\star) = 0$*). Assume in addition that for all* $o^{1\star} = (z^1, \omega^1)$*, the class of derivatives of* $p_\theta^\star(z^1|\omega^1)$ *with respect to* θ *is uniformly bounded (in* θ*) by an integrable function (of* z^1*). Then the covariance matrix* Σ *satisfies*

$$\Sigma = \frac{1}{2}{S'_{\theta_0}}^{-1} E_{P_1^\star}[\bar{q}_0(O^\star)^2 \dot{\ell}^\star_{\theta_0}(O^\star)\dot{\ell}^\star_{\theta_0}(O^\star)^\top]{S'_{\theta_0}}^{-1}.$$

In particular, 2Σ can be interpreted as the asymptotic covariance matrix of the M-estimator of θ_0 based on the loss function $\theta \mapsto \bar{q}_0(O^\star)\log p_\theta^\star(O^\star)$ and n iid observations drawn from $P_1^\star$. The n observations under P_0-case-control sampling correspond to $2 \times n$ observations under $P_1^\star$-prospective sampling, each of them in the former counting for two in the latter. Elements of proof are relegated to the appendix.

5 Threshold regression parametric modeling

5.1 Health as a stochastic process

We adopt the threshold regression approach (see [6, 7] and references therein), that is (quoting the title of [6]) we model the time to event of interest (development of an incident lung cancer) as a stochastic process reaching a boundary. The latter stochastic process represents here the amount of health relative to lung cancer. As long as it stays above zero (the so-called boundary), no lung cancer occurs. Crossing the boundary for the first time corresponds to developing an incident lung cancer.

Let $\mathbb{B}$ be a Brownian motion. For any real numbers $h > 0$ and $\mu \leq 0$, define

$$T(h, \mu) = \inf\{t \geq 0 : h + \mu t + \mathbb{B}_t \leq 0\}, \tag{14}$$

the first time the drifted Brownian motion $(h + \mu t + \mathbb{B}_t, t \geq 0)$ hits the set of nonnegative numbers. The distribution of $T(h, \mu)$ is known as the inverse Gaussian distribution with parameter (h, μ). It is characterized by its cumulative distribution function (cdf)

$$F(t; h, \mu) = 1 + e^{-2h\mu}\Phi\left((\mu t - h)t^{-1/2}\right) - \Phi\left((\mu t + h)t^{-1/2}\right),$$

where Φ is the standard normal cdf.

It is well known (see for instance [4]) that the drifted Brownian motion $(h+\mu t+\mathbb{B}_t, t \geq 0)$ will almost surely eventually reach the boundary (*i.e.*, $T(h, \mu) < \infty$) because $\mu \leq 0$. Therefore $T(h, \mu)$ is also characterized by its density

$$f(t; h, \mu) = \frac{h}{(2\pi t^3)^{1/2}} \exp\left(-\frac{(h - |\mu|t)^2}{2t}\right).$$

Finally, $T(h, \mu)$ has mean $h/|\mu|$ whenever $\mu < 0$.

Here, $(h + \mu t + \mathbb{B}_t, t \geq 0)$ models the amount of health relative to lung cancer as affected by the exposure to asbestos *in absence of such an exposure*, so that an incident lung cancer eventually occurs at time $T(h, \mu)$. Furthermore, this presentation of the model we are building yields a nice interpretation of the parameter (h, μ): h plays the role of an initial amount of health relative to lung cancer, and μ a rate of decay of the latter amount of health.

Describing what happens *in presence of exposures* involves the introduction of an acceleration function R, that is a nondecreasing continuous function on the nonnegative real line such that $R(t) \geq t$ for all t. Given such a function R, we define

$$T(h, \mu, R) = \inf\{t \geq 0 : h + \mu R(t) + \mathbb{B}_{R(t)} \leq 0\}, \tag{15}$$

the first time the drifted Brownian motion $(h + \mu t + \mathbb{B}_t, t \geq 0)$ hits the set of nonnegative numbers *along the modified time scale* derived from R. 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,

$$T(h, \mu, R) \geq t \quad \text{if and only if} \quad T(h, \mu) \geq R(t), \tag{16}$$

so that the cdf of $T(h, \mu, R)$ at t is $F(R(t); h, \mu)$, and its density at t is $R'(t)f(R(t); h, \mu)$ as soon as R is differentiable.

More importantly here, by virtue of the factorization of the likelihood exhibited in (3), 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 respectively

$$G(t; h, \mu, R) = \frac{1 - F(R(t); h, \mu)}{1 - F(R(x - 1); h, \mu)} \tag{17}$$

and

$$g(t; h, \mu, R) = \frac{R'(t)f(R(t); h, \mu)}{1 - F(R(x - 1); h, \mu)}. \tag{18}$$

The time $T(h, \mu, R)$ should be interpreted as the time to incident lung cancer under a history of exposure to asbestos *compatible* with R, a notion we investigate in the next section.

5.2 Calendar versus biological ages: modeling the ageing acceleration due to occupational exposure to asbestos

There is a nice interpretation of the acceleration function device. Admitting that the reference time scale (that is that of the Brownian motion $\mathbb{B}$) corresponds to *chronological/calendar* time scale, the new time scale formed by an acceleration function R may be understood as a *biological* time scale. This interpretation acknowledges the fact that the ageing phenomenon related to lung cancer is stronger in presence of noxious exposure than in its absence.

We present now an original class of acceleration functions tailored to our particular description of occupational exposures. Let us define

$$\mathcal{M} = \Big\{(M_0, (M_{k,l})_{k,l\leq 3}) \in \mathbb{R}_+ \times \mathbf{M}_{3,3}(\mathbb{R}_+) : \\ 0 \leq M_{k,1} \leq M_{k,2} \leq M_{k,3} = 1, k = 1, 2, 3\Big\}. \tag{19}$$

Then the rate yielded by description $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3) \in \mathcal{E} \setminus \{0\}$ for $M \in \mathcal{M}$ writes as

$$M(\varepsilon) = 1 + M_0 \times M_{1,\varepsilon_1} \times M_{2,\varepsilon_2} \times M_{3,\varepsilon_3}, \tag{20}$$

and that of $\varepsilon = 0$ is set to $M(0) = 1$. Notably, M_0 is to be interpreted as the factor of acceleration of time for the higher exposure, which we recall is encoded by $\varepsilon = (3, 3, 3)$. Rates $M(\varepsilon)$ range from 1 to $M(3, 3, 3) = 1 + M_0$ and (with convention $0/0 = 1$)

$$\frac{M(\varepsilon) - 1}{M_0} = M_{1,\varepsilon_1} \times M_{2,\varepsilon_2} \times M_{3,\varepsilon_3} :$$

an exposure characterized by "probability/frequency/intensity" description $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$ achieves a fraction $M_{1,\varepsilon_1} \times M_{2,\varepsilon_2} \times M_{3,\varepsilon_3}$ of the maximal acceleration.

Note that we only need 7 parameters in order to fully describe the 28 possibly different rates of acceleration. Furthermore, it is easily seen that this parametrization is identifiable: if $M, M' \in \mathcal{M}$ satisfy $M(\varepsilon) = M'(\varepsilon)$ for all $\varepsilon \in \mathcal{E}$ then $M = M'$.

Consider $M \in \mathcal{M}$ and a generic longitudinal description $\bar{\varepsilon}$ as presented in Section 2. Let us define now the function $\tilde{r}$ over the nonnegative real line such that for all $t \geq 0$,

$$\tilde{r}(t; M, \bar{\varepsilon}) = M(\varepsilon(t)).$$

Function $\tilde{r}$ is piecewise constant, but we can construct a continuous version r such that, if $\tilde{r} = \sum_{l=1}^{L} \rho_l \mathbf{1}]t_l; t_{l+1}] + \mathbf{1}]t_{L+1}; \infty)$, then $r(t_l) = \rho_l$ for $l = 1, \ldots, L$ and $\|\tilde{r} - r\|_\infty \leq \alpha$ for a small $\alpha > 0$ chosen a priori. Because we are willing to add a constraint $M(0) \leq C_1$ to the definition of $\mathcal{M}$ and to impose that a generic longitudinal description cannot have more than C_2 breakpoints (that is to upper bound a priori the number of employments that a subject can have in a lifetime), the mapping $\tilde{r} \mapsto r$ can even guarantee $\sup_{M,\bar{\varepsilon}} \|\tilde{r} - r\|_\infty \leq \alpha$. We will denote hereafter by $r(\cdot; M, \bar{\varepsilon})$ the continuous function associated to M and $\bar{\varepsilon}$ and proceed as if $\alpha = 0$.

Finally, every pair $(M, \bar{\varepsilon})$ gives rise to the *differentiable* (because $r(\cdot; M, \bar{\varepsilon})$ is continuous) acceleration function characterized by

$$R(t; M, \bar{\varepsilon}) = \int_0^t r(s; M, \bar{\varepsilon}) ds.$$

In particular if $\varepsilon(t) = 0$ for all $t \geq 0$ (that is in absence of exposure throughout lifetime), then $R(t; M, \bar{\varepsilon}) = t$ for all $t \geq 0$: in other words, the chronological and biological time scales coincide.

Now, given parameters $h > 0$, μ, $M \in \mathcal{M}$ and covariate $\bar{a}$, we obtain $R_a = R(\cdot; M, \bar{a})$, which yields in turn the time to incident lung cancer $T(h, \mu, R_a)$.

5.3 A notion of maximal number of years of life guaranteed free of lung cancer

Equivalence (16) has another important straightforward consequence:

$$T(h, \mu) = R(T(h, \mu, R)).$$

In particular, given parameters $h > 0$, μ, $M \in \mathcal{M}$ and covariate $\bar{a}$, $T(h, \mu) = R_a(T(h, \mu, R_a))$. In words, all things (gender, occurrence of lung cancer in close family, lifetime tobacco use) being equal, the age at incident lung cancer *in the absence of occupational exposure to asbestos* can be deduced deterministically from the (observed) age at incident lung cancer and history of occupational exposure to asbestos of a case. Furthermore, the nonnegative quantity

$$R_a(T(h, \mu, R_a)) - T(h, \mu, R_a)$$

(with convention $R_a(\infty) = \infty$ and $\infty - \infty = 0$) can be interpreted as a *maximal number of years of life guaranteed free of lung cancer*. The expression conveys the notion that $R_a(T(h, \mu, R_a)) - T(h, \mu, R_a)$ is different from the remaining number of years of life, as death may occur anytime after $T(h, \mu, R_a)$ even in the absence of occupational exposure to asbestos. Heuristically, it is the number of years of life which a subject living infinitely would enjoy before developing an incident lung cancer

5.4 Case-control weighted log-likelihood loss function

We derive in this section a valid log-likelihood loss function based on the threshold regression parametric modelling introduced in Sections 5.1 and 5.2. By Section 3, we know that it suffices to model the distribution of the observed data structure $O^\star$ under prospective sampling.

As explained in Section 3, we wish to model parametrically the conditional distribution of $O^\star$ given $\Omega = (W, X, \bar{A}(X), Y)$, leaving the conditional distribution of Ω given Y unspecified.

For this purpose, we state that under $\theta = (\alpha, \beta, M) \in \Theta = \mathbb{R}^4 \times \mathbb{R}^{16} \times \mathcal{M}$, the conditional distribution of T (the possibly unobserved age at incident lung cancer of the subject associated with $O^\star$) given Ω is that of $T(h, \mu, R_a)$ with

$$\log h = \alpha_{1+(0,1,2,0)W^\top} \tag{21}$$

(each level of $(W_1, W_2) \in \{0,1\}^2$ is associated with a unique positive initial health h),

$$\log(-\mu) = \beta_{1+(0,1,2,4)W^\top} \tag{22}$$

(each level of $(W_1, W_2, W_3) \in \mathcal{W} = \{0,1\}^2 \times \{0,1,2,3\}$ is associated with a unique negative drift μ), and

$$R_a = R(\cdot; M, \bar{A}(X)).$$

Therefore,

$$\log p^\star_\theta(Z|\Omega) = Y \log g(Z;\theta) + (1-Y) \log G(Z;\theta)$$

with convention

$$\begin{aligned} G(Z;\theta) &= G(Z; h, \mu, R(\cdot; M, \bar{A}(X))), \\ g(Z;\theta) &= g(Z; h, \mu, R(\cdot; M, \bar{A}(X))), \end{aligned}$$

the functions G and g being defined in (17) and (18). Finally, the relevant part of the resulting case-control weighted log-likelihood at $\theta \in \Theta$ writes as

$$\sum_{i=1}^{n} \left\{ q_0 \log g(Z_i^1;\theta) + \bar{q}_0(V_i^1) \frac{1}{J_i} \sum_{j=1}^{J_i} \log G(Z_i^{0,j};\theta) \right\} = P_n \tilde{\ell}(\cdot|\theta), \tag{23}$$

where $\tilde{\ell}(O|\theta) = \ell(O|p^\star_\theta) - \mathrm{rem}(O)$ (see equation (10)) and $P_n = \sum_{i=1}^n \delta_{O_i}$ denotes the empirical measure.

5.5 Multi-fold likelihood-based cross-validation

It makes no doubt that the model we have built so far is over-dimensioned. The "probability/frequency/intensity" description with its 28 different levels is itself certainly too rich (see again Table 1), or at least difficult to establish and prone to errors. We rather consider the model Θ described so far as a "maximal" model giving rise to a large collection of sub-models Θ_k obtained by adding constraints on the "maximal" parameter $\theta = (\alpha, \beta, M) \in \Theta$. The number of such sub-models is large indeed: there are $(1 + 7^3) = 344$ sub-models 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 sub-models equals $2^2 \times 2^3 \times 344 = 11,008$. It is out of question to explore the whole collection of sub-models. Instead, we propose to

(i) define a large collection $\{\Theta_k : k \in \mathcal{K}\}$ of sub-models of interest,

(ii) let the data select a better sub-model $\Theta_{\hat{k}}$ in the latter collection based on a multi-fold likelihood-based cross-validation criterion.

It is shown in [15] 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 the likelihood risk of the better model will not be much bigger than that of the oracle model. Although we cannot invoke rigorously this remarkable property here, it motivates the procedure that we describe below.

The likelihood risk of $\theta \in \Theta$ is by definition

$$\mathcal{R}(\theta) = -E_{P_0}\tilde{\ell}(O|\theta),$$

which is closely related to minus the Kullback-Leibler divergence between the density $p_0^\star$ of $P_0^\star$ and $p_\theta^\star$, as explained in Section 3. Let us denote by $\theta_{n,k}(P_n)$ the case-control weighted maximum likelihood estimator defined in (9) with θ ranging over Θ_k. Given the collection $\{\theta_{n,k}(P_n) : k \in \mathcal{K}\}$ we wish to select the estimator $\theta_{n,\bar{k}}(P_n)$ that minimizes $\mathcal{R}$, where $\bar{k}$ itself depends on P_n. Because the definition of $\mathcal{R}$ involves the true distribution P_0, we must estimate $\mathcal{R}(\theta_{n,k}(P_n))$ and choose to do so by multi-fold cross-validation. Details follow.

We split the data randomly into a *training* and a *validation* samples. Given an integer V (later set to $V = 10$), each observed data structure O_i is associated with a label $\text{lab}_i = 1 + (i \mod V)$. The collection of labels $\{\text{lab}_i : i \leq n\} \subset \{1, \ldots, V\}$ is such that $\max_{l,l' \leq V} |\sum_{i=1}^n \mathbf{1}\{\text{lab}_i = l\} - \sum_{i=1}^n \mathbf{1}\{\text{lab}_i = l'\}| \leq 1$. The splitting random variable $S = (S_1, \ldots, S_n) \in \{0,1\}^n$ is drawn independently of $O_1, \ldots, O_n$ in such a way that, for each $1 \leq l \leq V$, $S = (\mathbf{1}\{\text{lab}_1 = l\}, \ldots, \mathbf{1}\{\text{lab}_n = l\})$ with probability V^{-1}. Conditionally on S, the observed data structure O_i belongs to the training sample if $S_i = 0$ (there are approximately $n(V-1)/V$ such O_i's), otherwise it belongs to the validation sample. The empirical distribution of those O_i's for which $S_i = 0$ (respectively, $S_i = 1$) is $P_{n,S}^0$ (respectively, $P_{n,S}^1$). The empirical distribution of those O_i's for which $\text{lab}_i = l$ (respectively, $\text{lab}_i \neq l$) is P_n^l (respectively, P_n^{-l}).

Each Θ_k yields a maximum likelihood estimator $\theta_{n,k}(P_{n,S}^0)$ based on the training sample only. Its risk, averaged over the splits, writes as

$$\text{crit}(k) = E_S \mathcal{R}(\theta_{n,k}(P_{n,S}^0)) = -\frac{1}{V}\sum_{l=1}^V E_{P_0}\tilde{\ell}\left(O|\theta_{n,k}(P_n^{-l})\right).$$

The value $\tilde{k}$ that minimizes $k \mapsto \text{crit}(k)$ over $\mathcal{K}$ is called the *oracle* because it depends both on P_n and on P_0. In our attempt to reach that $\tilde{k}$ which is a good proxy to $\bar{k}$, we estimate $\text{crit}(k)$ by

$$\widehat{\text{crit}}(k) = -E_S E_{P_{n,S}^1}\tilde{\ell}\left(O|\theta_{n,k}(P_{n,S}^0)\right) = -\frac{1}{V}\sum_{l=1}^V E_{P_n^l}\tilde{\ell}\left(O|\theta_{n,k}(P_n^{-l})\right),$$

and propose to use the value $\hat{k}$ that minimizes $k \mapsto \widehat{\text{crit}}(k)$ over $\mathcal{K}$, whose definition is postponed to Section 6.2. In conclusion, the final estimator is $\theta_{n,\hat{k}}(P_n)$.

$q_0 = 470.0682\text{e-}06$

a	$q_0(1\|0,a)$	$q_0(1\|1,a)$
1	2.058932e-06	1.663324e-06
2	1.859944e-05	1.460473e-05
3	6.803086e-05	4.461827e-05
4	2.586692e-04	1.184914e-04
5	6.484864e-04	1.947058e-04
6	1.192778e-03	2.542976e-04
7	1.854668e-03	3.294062e-04
8	2.331553e-03	3.588764e-04
9	2.928415e-03	4.466062e-04
10	3.686216e-03	5.312313e-04
11	3.608302e-03	5.332930e-04
12	3.636995e-03	5.395069e-04
13	2.171286e-03	3.234775e-04

a	$\bar{q}_0(0,a)$	$\bar{q}_0(1,a)$
1	228.3063171	282.6071669
2	25.2727716	32.1855444
3	6.9091613	10.5348607
4	1.8167860	3.9666376
5	0.7243998	2.4137787
6	0.3936251	1.8480261
7	0.2529813	1.4265470
8	0.2011416	1.3093632
9	0.1600496	1.0520638
10	0.1270504	0.8843954
11	0.1298040	0.8809745
12	0.1287763	0.8708223
13	0.2160229	1.4527010

Table 3: Estimating the probability distribution of being a case, based on the independent study [2]. Left: Estimates of $q_0(1|w_1, v_2)$, as defined in (5). Middle: Estimate of q_0, as defined in (4). Right: Estimates of $\bar{q}_0(w_1, v_2)$, as defined in (7). Here, $w_1 = 0$ for men and $w_1 = 1$ for women, and $v_2 = a$ if the age at sampling x belongs to $[t_a; t_{a+1})$, where $t_0 = 0$, $t_a = 30 + 5(a-1)$ for $1 \leq a \leq 12$ and $t_{13} = \infty$.

6 Results

6.1 Conditional distribution of being a case

Estimating the marginal probability of being a case q_0 (4), the conditional probabilities of being a case or a control conditional on the matching variable $(q_0(1|v))_{v\in\mathcal{V}}$ (5), and the weights $(\bar{q}_0(v))_{v\in\mathcal{V}}$ (7) is made possible thanks to [2], an independent study of cancer incidence and mortality in France, over the period 1980–2005. However, we must assume either *(i)* that the data from [2], which are collected over the whole French population, are representative of the Parisian population of interest, or *(ii)* as underlined in Section 2 that sampling from the four Parisian hospitals that participate to the study is stochastically equivalent to sampling from the population of France.

We first estimate these quantities for each year from 1999 to 2002 separately. In agreement with our stationary assumption (1), we remark that the various estimates are very consistent over the years. In order to gain precision, we average the estimates over the years. The final estimates are presented in Table 3. We emphasize that the weights $(\bar{q}_0(v))_{v\in\mathcal{V}}$ are far from being homogeneous.

6.2 Model selection procedure in action

We explain in Section 5.5 how the best model index $\hat{k}$ (with related best model $\Theta_{\hat{k}}$) is obtained in a pre-determined collection $\mathcal{K}$ of sub-model indices (with related sub-models Θ_k, $k \in \mathcal{K}$). The latter collections are constructed by recursion as presented below.

We first initialize $\Theta^0 = \Theta$ and $\mathcal{K}^{-1} = \emptyset$ with convention $\max \emptyset = 0$.

At a given step $\nu \geq 0$, a sub-model Θ^ν is defined as a subset of Θ meeting ν independent

one-dimensional constraints on $M \in \mathcal{M}$ (*i.e.*, constraints of the type $M_{k,l-1} = M_{k,l}$ for some $k = 1, 2, 3$ and $l = 1, 2, 3$ with convention $M_{k,0} = 0$). Start with $c(\nu+1) = -\infty$ and $\Theta^{\nu+1} = \emptyset$. The following rule is applied to $\Theta^{\nu+1}$:

Rule 1. For every possible sub-model $\Theta' \subset \Theta^\nu$ derived from Θ^ν by adding another one-dimensional constraint on M as described above (all such models share the same dimension), evaluate the corresponding maximum log-likelihood criterion

$$\ell(\Theta') = \max_{\theta \in \Theta'} P_n \tilde{\ell}(\cdot|\theta).$$

If $\ell(\Theta') \geq c(\nu + 1)$, update $c(\nu + 1) = \ell(\Theta')$ and $\Theta^{\nu+1} = \Theta'$.

Applying Rule 1 as long as possible yields 7 sets Θ^ν, $\nu = 0, \ldots, 6$. Their description is given in Table 4.

$$\begin{aligned}
\Theta^0 &= \Theta, \\
\Theta^1 &= \{\theta \in \Theta^0 : M_{1,1} = 0\}, && (24)\\
\Theta^2 &= \{\theta \in \Theta^1 : M_{2,1} = M_{2,2}\}, && (25)\\
\Theta^3 &= \{\theta \in \Theta^2 : M_{2,2} = 1\}, && (26)\\
\Theta^4 &= \{\theta \in \Theta^3 : M_{3,1} = M_{3,2}\}, && (27)\\
\Theta^5 &= \{\theta \in \Theta^4 : M_{1,2} = 1\}, && (28)\\
\Theta^6 &= \{\theta \in \Theta^5 : M_{3,1} = 0\}. && (29)
\end{aligned}$$

(24) low probability does not differ from no exposure at all;
(25) moreover, low and mild frequencies do not differ;
(26) moreover, mild and high frequencies do not differ;
(27) moreover, low and mild intensities do not differ;
(28) moreover, mild and high probabilities do not differ;
(29) moreover, low intensity does not differ from no exposure.

Table 4: Descriptions of $\Theta^0, \ldots, \Theta^6$. The collection of parameter sets is nested. For instance, Θ^3 is the set of those $\theta \in \Theta$ such that $M_{1,1} = 0$ and $M_{2,1} = M_{2,2} = 1$. Regarding dimensions, it trivially holds that, for each $0 \leq k \leq 6$, $\dim(\Theta^k) = 27 - k$.

At a given step $\nu \geq 0$, a set $\mathcal{K}^{\nu-1}$ of successive integers is defined. Start with $\mathcal{K}^\nu = \{\max K^{\nu-1} + 1\}$ (a set initially containing a single element) and define $\Theta^{\nu,\max K^{\nu-1}+1} = \Theta^\nu$. The following second rule is applied to $\mathcal{K}^\nu$:

Rule 2. For every possible constraint "$\varphi(\theta) = 0$" on $\theta \in \Theta^\nu$ of the form "α and β independent of W_l" for some $l = 1, 2, 3$ (the lth coordinate of W does not affect the value of the initial health and drift parameters h and μ, see (21) and (22)), update $\mathcal{K}^\nu = \mathcal{K}^\nu \cup \{\max K^\nu + 1\}$ and define $\Theta^{\nu,\max K^\nu} = \{\theta \in \Theta^\nu : \varphi(\theta) = 0\}$.

(Note that each Θ^ν therefore gives rise to $2^3 = 8$ sub-models $\Theta^{\nu,l}$.)

We apply Rule 2 for $\nu = 0, \ldots, 6$, and finally define

$$\mathcal{K} = \cup_{\nu=0}^{11} \mathcal{K}^\nu = \{1, 2, 3, \ldots, 56\}.$$

For every $k \in \mathcal{K}$ there exists a unique $\nu = 0, \ldots, 7$ such that $\Theta^{\nu,k}$ be defined: setting $\Theta_k = \Theta^{\nu,k}$ concludes the definition of the collection $\{\Theta_k : k \in \mathcal{K}\}$ of sub-models of interest.

The best model $\Theta_{\hat{k}}$ (according to our multi-fold likelihood-based cross-validation criterion) is a subset of Θ^2, featuring 16 degrees of freedom. Its complete description follows:

- the *initial health* parameter depends on W only through gender (hence not on the indicator of occurrence of lung cancer in close family);
- the *drift* parameter depends on W only through gender and lifetime tobacco use (hence not on the indicator of occurrence of lung cancer in close family);
- *exposure to asbestos* is significantly noxious; there is no difference between low probability and no exposure to asbestos at all (in view of (19), $M_{1,1} = 0$) and no difference either between low and mild frequencies (in terms of (19), $M_{2,1} = M_{2,2}$).

6.3 Fitting the best model

The best model $\Theta_{\hat{k}} \subset \Theta^2$ described in Section 6.2 is first fitted in terms of maximum likelihood on the whole dataset. Regarding the derivation of confidence intervals, we decide to rely on the bootstrap instead of a central limit theorem (such as Proposition 3). The particulars of the bootstrap procedure follow. We set $\alpha = 2.5\%$, $B = 1,000$ 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 $P^b_{n(1-p)}$, in order to compute and store the corresponding maximum likelihood estimate $\theta_{n(1-p),\hat{k}}(P^b_{n(1-p)})$ of $\theta \in \Theta_{\hat{k}}$. The mean and median values of $\boldsymbol{\theta}^B_{n,\hat{k}} = \{\theta_{n(1-p),\hat{k}}(P^b_{n(1-p)}) : b \leq B\}$ only very slightly differ from each other (moreover, they are very close to the maximum likelihood estimate $\theta_{n,\hat{k}}(P_n)$ computed on the whole dataset). The componentwise $\alpha/16$- and $(1-\alpha/16)$-quantiles of $\boldsymbol{\theta}^B_{n,\hat{k}}$ are used as lower- and upper-bounds of confidence intervals, which simultaneously provide a $(1-2\alpha) = 95\%$-coverage by the applied Bonferroni correction. Specifically:

- *initial health*:

W_1	h
0	23.82 [23.42; 24.13]
1	25.09 [24.86; 25.40]

  It is seen in particular that women are associated with a significantly larger initial health than men.

- *drift*:

	-100μ	
W_3	$W_1 = 0$	$W_1 = 1$
0	0.69 [0.08; 1.46]	0.02 [0.01; 0.03]
1	7.70 [6.91; 8.28]	6.63 [5.73; 7.68]
2	13.89 [13.25; 14.46]	10.55 [9.63; 11.80]
3	17.67 [17.11; 18.38]	14.79 [13.65; 17.77]

  Two main features arise:

– For each level of lifetime tobacco use, the absolute value of the drift is significantly larger for men than for women (actually, the confidence intervals for $W_3 = 3$ slightly overlap). Combined with the already mentioned fact that women are associated with a larger initial health, this implies that *for any given history of exposure to asbestos* and for every level of lifetime tobacco use, the distribution of age at incident lung cancer in women is stochastically dominated by the distribution of age at 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, it is argued in [16] that women are more susceptible to tobacco carcinogens, but it is seen in [5] that men *or* women are more susceptible to tobacco carcinogens, depending on ethnic and racial group).

– Both in men and women, the absolute value of the drift 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 age at incident lung cancer for lifetime tobacco use equal to w is stochastically dominated by the distribution of age at 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 [3].

- *exposure to asbestos*:

M_0: 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 $[0.00; 0.27]$	$M_{3,3} = 1$

  We notably derive from the above table the values of $(M(\varepsilon) - 1)$ (which can be interpreted as a factor of acceleration of time due to an exposure of level ε, see (20)) and related confidence intervals for each level of exposure $\varepsilon \in \mathcal{E} \setminus \{0\}$, see Table 5.

6.4 Application to the maximal number of years of life guaranteed free of lung cancer

In view of Section 5.3 and the notion of maximal number of years of life guaranteed free of lung cancer, the results of the previous section provide us with a way of evaluating the latter number on a case by case basis. Arguably, we mostly care for a pointwise estimation of, and confidence lower-bound on, the maximal number of years of life guaranteed free of lung cancer. In order to address this issue, let us compute a counterpart of Table 5 based on the componentwise $2\alpha/5$-quantiles of $\boldsymbol{\theta}^B_{n,\hat{k}}$ (which simultaneously provide $(1-2\alpha) = 95\%$-coverage for parameter M *on its own* by the applied Bonferroni correction, since there M has 5 degrees of freedom), see Table 6.

ε	$M(\varepsilon)-1$	ε	$M(\varepsilon)-1$	ε	$M(\varepsilon)-1$
111	0	211	0.026 $[0.000; 0.171]$	311	0.026 $[0.000; 0.173]$
112	0	212	0.092 $[0.001; 0.530]$	312	0.094 $[0.001; 0.537]$
113	0	213	1.078 $[0.297; 1.939]$	313	1.108 $[0.309; 1.964]$
121	0	221	0.026 $[0.000; 0.171]$	321	0.026 $[0.000; 0.173]$
122	0	222	0.092 $[0.001; 0.530]$	322	0.094 $[0.001; 0.537]$
123	0	223	1.078 $[0.297; 1.939]$	323	1.108 $[0.309; 1.964]$
131	0	231	0.027 $[0.000; 0.174]$	331	0.028 $[0.000; 0.176]$
132	0	232	0.099 $[0.001; 0.539]$	332	0.101 $[0.001; 0.546]$
133	0	233	1.159 $[0.330; 1.971]$	333	1.192 $[0.344; 1.998]$

Table 5: Estimated values (with precision 10^{-3}) of the factor of acceleration of time ($M(\varepsilon)-1$) and related confidence intervals for each level of exposure $\varepsilon \in \mathcal{E} \setminus \{0\}$. Recall that $M(0) = 1$.

ε	$M(\varepsilon)-1$	ε	$M(\varepsilon)-1$	ε	$M(\varepsilon)-1$
111	0	211	0.026 $[0.001; \infty)$	311	0.026 $[0.001; \infty)$
112	0	212	0.092 $[0.004; \infty)$	312	0.094 $[0.004; \infty)$
113	0	213	1.078 $[0.374; \infty)$	313	1.108 $[0.389; \infty)$
121	0	221	0.026 $[0.001; \infty)$	321	0.026 $[0.001; \infty)$
122	0	222	0.092 $[0.004; \infty)$	322	0.094 $[0.004; \infty)$
123	0	223	1.078 $[0.374; \infty)$	323	1.108 $[0.389; \infty)$
131	0	231	0.027 $[0.001; \infty)$	331	0.028 $[0.002; \infty)$
132	0	232	0.099 $[0.004; \infty)$	332	0.101 $[0.004; \infty)$
133	0	233	1.159 $[0.414; \infty)$	333	1.192 $[0.431; \infty)$

Table 6: Estimated values (with precision 10^{-3}) of the factor of acceleration of time ($M(\varepsilon)-1$) and related *right* confidence intervals for each level of exposure $\varepsilon \in \mathcal{E} \setminus \{0\}$. Recall that $M(0) = 1$. A Bonferroni correction ensures that the confidence regions simultaneously guarantee $(1-2\alpha) = 95\%$-coverage (for $\{M(\varepsilon) - 1 : \varepsilon \in \mathcal{E}\}$ on its own).

Elementary algebra permits to compute an evaluation $c(t, \bar{a}(t))$ of, and confidence lower-bound $c^-(t, \bar{a}(t))$ on, the maximal number of years of life guaranteed free of lung cancer for any couple $(t, \bar{a}(t))$ of age t at incident lung cancer and history $\bar{a}(t)$ of occupational exposure to asbestos till t. Let us consider three examples:

- Consider a case of incident lung cancer at age t who spent, till that age, 30 years with an occupational exposure to asbestos $\varepsilon = 332$: one evaluates $c(t, \bar{a}(t)) = 30 \times 0.101 = 3.03$ maximal number of years guaranteed free of lung cancer, with its 95%-confidence lower bound $c^-(t, \bar{a}(t)) = 30 \times 0.004 = 0.09$ maximal number of years guaranteed free of lung cancer (approximately 44 days).

  This is quite an extreme case, since 3 out of the 8,432 employments described in the dataset achieve the description $\varepsilon = 332$.

- Consider a case of incident lung cancer at age t who spent, till that age, 10 years (then later 5 years and 2 years) with an occupational exposure to asbestos $\varepsilon = 322$ (then later $\varepsilon = 121$ and $\varepsilon = 222$): one evaluates $c(t, \bar{a}(t)) = 10 \times 0.094 + 5 \times 0 + 2 \times 0.092 = 1.124$ maximal number of years guaranteed free of lung cancer, with its 95%-confidence lower bound $c^-(t, \bar{a}(t)) = 10 \times 0.004 + 5 \times 0 + 2 \times 0.004 = 0.048$ maximal number of years guaranteed free of lung cancer (approximately 17.5 days).

  Note that 150, 36 and 189 out of the 8,432 employments described in the dataset achieve the descriptions $\varepsilon = 121$, $\varepsilon = 222$ and $\varepsilon = 322$.

- Consider a case of incident lung cancer at age t who spent, till that age, 10 years (then later 15 years) with an occupational exposure to asbestos $\varepsilon = 213$ (then later $\varepsilon = 223$): one evaluates $c(t, \bar{a}(t)) = 10 \times 1.078 + 15 \times 1.078 = 26.95$ maximal number of years guaranteed free of lung cancer, with its 95%-confidence lower bound $c^-(t, \bar{a}(t)) = 10 \times 0.374 + 15 \times 0.374 = 9.350$ maximal number of years guaranteed free of lung cancer.

  This is quite an extreme case, since only 6 and 3 out of the 8,432 employments described in the dataset achieve the descriptions $\varepsilon = 213$ and $\varepsilon = 223$.

Among the $n = 860$ cases of our dataset, only 259 (*i.e.*, 30%) cases are associated with positive maximal number of years of life guaranteed free of lung cancer. We report in Table 7 the quartiles, mean and extreme values of maximal number of years of life guaranteed free of lung cancer as computed on those 259 cases.

- The maximum value is reached by a male who accumulated through his professional life a total of 33 years with occupational exposure to asbestos equal to $\varepsilon = 313$ and was diagnosed a lung cancer at 70 years old. Although this is not relevant as far as the evaluation of the potential years of life free of lung cancer is concerned, his lifetime tobacco equals 45 pack years.

- The minimum value is reached by 4 women who accumulated through their professional lives a total of 1 year with occupational exposure to asbestos $\varepsilon \in \{211, 221\}$ and were diagnosed a lung cancer at 51 (for two of them), 59 and 68 years old. Although this is not relevant as far as the evaluation of the potential years of life free of lung cancer is concerned, their lifetime tobacco uses equal 25, 30, 32 and 55 pack years).

- The median value is reached by a man who accumulated through his professional life a total of 4 years (respectively, 5 and 7) with occupational exposure to asbestos equal

	min.	25%	50%	mean	75%	max.
max. number of years free of lung cancer	0.026	0.289	0.769	2.467	2.408	36.577
95%-lower bound	0.001	0.014	0.037	0.555	0.102	12.832

Table 7: Quartiles, mean and extreme values of the maximal number of years of life guaranteed free of lung cancer and corresponding 95%-confidence lower-bound (with precision 10^{-3}), as computed on those 259 cases (*i.e.*, 30% of all cases) for whom the evaluated maximal number of years of life guaranteed free of lung cancer is positive.

to $\varepsilon = 111$ (respectively, $\varepsilon = 211$ and $\varepsilon = 212$) and was diagnosed a lung cancer at 71 years old. Although this is not relevant as far as the evaluation of the potential years of life free of lung cancer is concerned, his lifetime tobacco equals 55 pack years.

We represent in Figure 1 the empirical cdf of the maximal number of years of life guaranteed free of lung cancer (and corresponding 95%-confidence lower bounds) for the 259 cases for whom it is positive.

7 Discussion

We have developed a collection of threshold regression models (see Section 5), and have data-adaptively selected a better model in it by relying on multi-fold likelihood-based cross-validation (see Section 6.2 for the descriptions of the model selection procedure and derived better model). The latter better threshold regression model has been fitted by maximum likelihood, and bootstrapped confidence intervals have been obtained (see Section 6.3). The statistical procedure has been adjusted in order to eliminate the bias induced by the case-control sampling design used to collect the dataset. This necessary preliminary step was made possible because the probability distribution of being a case in the population of interest can be computed beforehand based on an independent study (see Section 6.1). We have discussed the implications of the fitted threshold regression model in terms of the notion of maximal number of years of life guaranteed free of lung cancer which is naturally attached to it (see Section 6.4).

We believe that, even though they cannot be interpreted causally, the results presented in this article contribute significantly to the quantitative understanding of how an occupational exposure to asbestos is related to an increase of lung cancer, and to the evaluation, in subjects suffering from a lung cancer, of how much the amount of exposure to asbestos explains the occurrence of the cancer.

We finally acknowledge a limitation of the approach undertaken in this article: The link between the occupational exposure to asbestos and age at incident lung cancer is well-defined in the context of the proposed threshold regression models, but we do not extend it beyond. The parameter we aim for is therefore difficult to comprehend (it is related to the Kullback-Leibler projection of the true distribution of the data onto a threshold regression model), and the inference procedure certainly fails to estimate optimally/efficiently what we really care for, which would be a measure of the strength of the link between the occupational exposure to asbestos and age at incident lung cancer defined non- or semiparametrically. We intend to go further in that direction in future work.

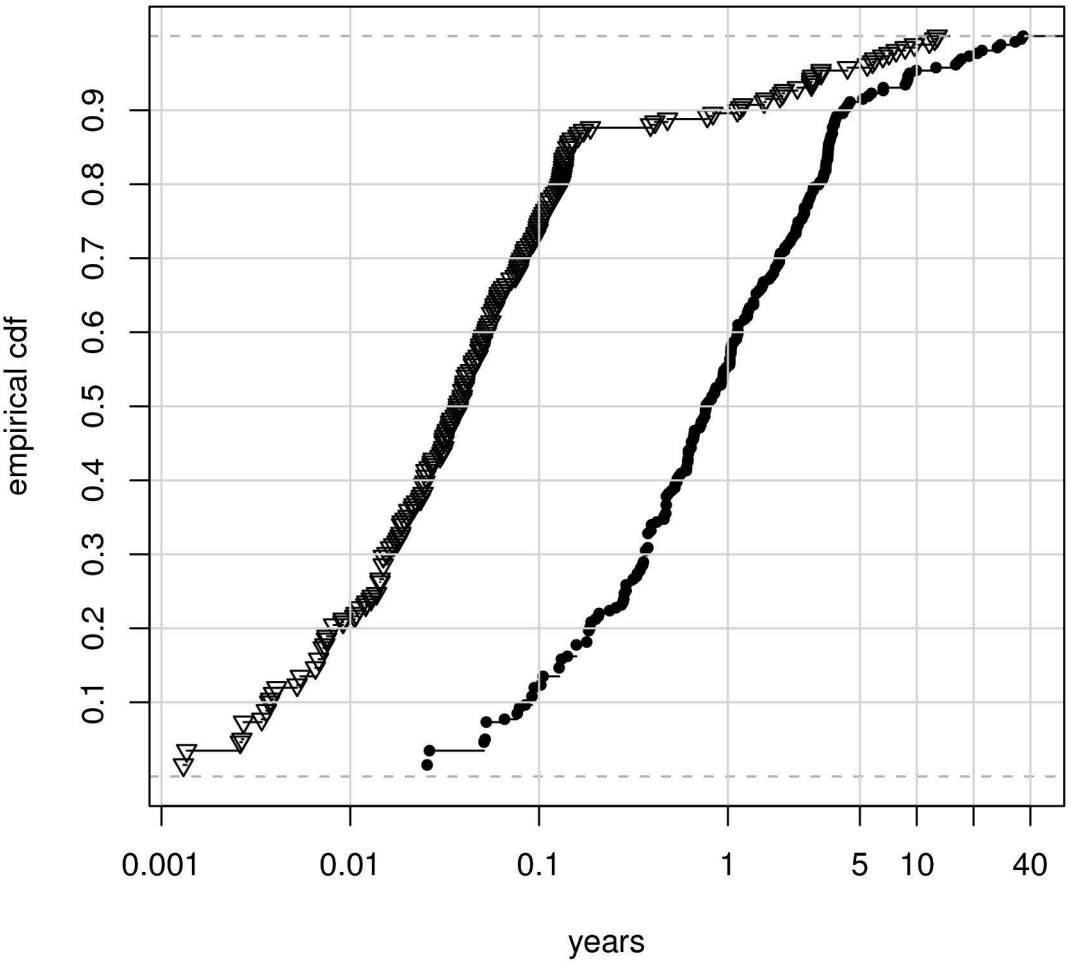

Figure 1: **Empirical distributions of maximal number of years of life guaranteed free of lung cancer and related confidence lower-bound.** The rightmost curve with bullets (respectively leftmost curve with triangles) represents the empirical cdf of the maximal number of years guaranteed free free of lung cancer (respectively of the 95%-confidence lower bound on that number) of those cases for whom it is positive, that is the empirical cdf of $\{c(T_i^1, \bar{A}^1(T_i^1)) : c(T_i^1, \bar{A}^1(T_i^1)) > 0, i \leq n\}$ (respectively $\{c^-(T_i^1, \bar{A}^1(T_i^1)) : c(T_i^1, \bar{A}^1(T_i^1)) > 0, i \leq n\}$). Only 30% of the cases are concerned. The x-axis scale is logarithmic.

A Appendix: elements of proof

Proof of Proposition 1. On one hand, note that

$$\begin{aligned} E_{P_0} q_0 \log p^\star(V^1, O^{1\star}) &= \int q_0 \log p^\star(v^1, o^{1\star}) dP_0^\star(v^1, o^{1\star}|y=1) \\ &= \int \log p^\star(v^1, o^{1\star}) dP_0^\star(v^1, o^{1\star}, y=1) \\ &= \int \log p^\star(o^\star) dP_0^\star(o^\star, y=1). \end{aligned} \tag{30}$$

On the other hand, for each $j \leq J$,

$$\begin{aligned} E_{P_0} \bar{q}_0(V^1) \log p^\star(V^1, O^{0,j\star}) &= E_{P_0} \bar{q}_0(V^1) E_{P_0}\left[\log p^\star(V^1, O^{0,j\star})|V^1\right] \\ &= E_{P_0} \bar{q}_0(V^1) \int \log p^\star(V^1, o^\star) dP_0^\star(o^\star|V^1, y=0) \\ &= \int \bar{q}_0(v^1) \log p^\star(v^1, o^\star) dP_0^\star(o^\star|v^1, y=0) dP_0(v^1). \end{aligned}$$

Furthermore, for each $v \in \mathcal{V}$, $dP_0(v) = dP_0^\star(v|y=1) = q_0(v|1)\delta_v(v)$ (we use the same shorthand notation as in (4), (5), (6), (7)) and denote by δ_v the Dirac mass at v), hence

$$\bar{q}_0(v) dP_0(v) = q_0 \frac{q_0(0|v)}{q_0(1|v)} q_0(v|1)\delta_v = q_0(0|v) P_0^\star(v)\delta_v(v) = dP_0^\star(v, y=0).$$

Consequently, we obtain

$$\begin{aligned} E_{P_0} \bar{q}_0(V^1) \log p^\star(V^1, O^{0,j\star}) &= \int \log p^\star(v^1, o^\star) dP_0^\star(o^\star|v^1, y=0) dP_0^\star(v^1, y=0) \\ &= \int \log p^\star(v^1, o^\star) dP_0^\star(v^1, o^\star, y=0) \\ &= \int \log p^\star(o^\star) dP_0^\star(o^\star, y=0) \end{aligned} \tag{31}$$

(which does not depend on j). Combining (30), (31) finally yields

$$E_{P_0} \ell(O|p^\star) = \int \log p^\star(o^\star) dP_0^\star(o^\star) = E_{P_0^\star} \log p^\star(O^\star).$$

The conclusion is straightforward, because

$$E_{P_0^\star} \log p^\star(O^\star) - E_{P_0^\star} \log p_0^\star(O^\star) = -\mathrm{KL}(p_0^\star, p^\star),$$

the opposite of the Kullback-Leibler divergence between $p_0^\star$ and $p^\star$, which is positive for $p^\star \neq p_0^\star$ and equals zero otherwise. □

Proof of Proposition 3. The expansion (11) and the related distributional limit result are a consequence of Theorem 5.23 in [14]. The fact that $S_{\theta_0} = E_{P_0^\star} \ddot{\ell}_{\theta_0}^\star(O^\star)$ is obtained by adapting

slightly the proof of Proposition 1. Regarding $E_{P_0}[\dot{\tilde{\ell}}(O|\theta_0)\dot{\tilde{\ell}}(O|\theta_0)^\top]$, let us abbreviate $xx^\top$ to x^2 and note that

$$\begin{aligned}\dot{\tilde{\ell}}(O|\theta_0)\dot{\tilde{\ell}}(O|\theta_0)^\top = &\Big[q_0\bar{q}_0(O^{1\star})\dot{\ell}^\star_{\theta_0}(V^1,O^{1\star})^2 + \bar{q}_0(V^1)\Big(\tfrac{1}{J}\textstyle\sum_j \bar{q}_0(O^{0,j\star})\dot{\ell}^\star_{\theta_0}(V^1,O^{0,j\star})\Big)^2\Big]\\ &+\Big[q_0\bar{q}_0(V^1)\dot{\ell}^\star_{\theta_0}(V^1,O^{1\star})\Big(\tfrac{1}{J}\textstyle\sum_j \bar{q}_0(O^{0,j\star})\dot{\ell}^\star_{\theta_0}(V^1,O^{0,j\star})\Big)^\top\\ &\qquad + q_0\bar{q}_0(V^1)\Big(\tfrac{1}{J}\textstyle\sum_j \bar{q}_0(O^{0,j\star})\dot{\ell}^\star_{\theta_0}(O^{0,j\star})\Big)\dot{\ell}^\star_{\theta_0}(O^{1\star})^\top\Big].\end{aligned}$$

The P_0-expected value of the first term between brackets is $E_{P_0^\star}\ddot{\ell}^\star_{\theta_0}(O^\star)$, as another simple adaptation of the proof of Proposition 1 straightforwardly yields. Moreover,

$$\begin{aligned}&E_{P_0}q_0\bar{q}_0(V^1)\dot{\ell}^\star_{\theta_0}(V^1,O^{1\star})\Big(\tfrac{1}{J}\textstyle\sum_j \bar{q}_0(O^{0,j\star})\dot{\ell}^\star_{\theta_0}(V^1,O^{0,j\star})\Big)^\top\\ &\quad= E_{P_0}\left[q_0\bar{q}_0(V^1)E_{P_0}\left(\dot{\ell}^\star_{\theta_0}(V^1,O^{1\star})\Big(\tfrac{1}{J}\textstyle\sum_j \bar{q}_0(O^{0,j\star})\dot{\ell}^\star_{\theta_0}(V^1,O^{0,j\star})\Big)^\top\middle| V^1\right)\right]\\ &\quad= E_{P_0}\left[q_0\bar{q}_0(V^1)E_{P_0}\left(\dot{\ell}^\star_{\theta_0}(V^1,O^{1\star})\middle| V^1\right)\right.\\ &\qquad\qquad\left.E_{P_0}\left(\Big(\tfrac{1}{J}\textstyle\sum_j \bar{q}_0(O^{0,j\star})\dot{\ell}^\star_{\theta_0}(V^1,O^{0,j\star})\Big)^\top\middle| V^1\right)\right]\end{aligned}$$

by conditional independence. Denote by $\Pi = E_{P_0}(\dot{\ell}^\star_{\theta_0}(V^1,O^{1\star})|O^{1\star}\setminus Z^1)$ the conditional expectation of $\dot{\ell}^\star_{\theta_0}(V^1,O^{1\star})$ given every component of $O^{1\star}$ but Z^1, that is given Ω^1 (compatible with V^1). The projection Π can be written as a measurable function of Ω^1 times

$$\int \dot{\ell}^\star_{\theta_0}(z,\Omega^1)p^\star_{\theta_0}(z|\Omega^1)dz = \int \left.\frac{\partial p^\star_\theta(z|\Omega^1)}{\partial\theta}\right|_{\theta=\theta_0} dz = 0,$$

provided that the order of differentiation and integration can be reversed. This is ensured by the stated constraint on the derivatives of $p^\star_\theta(z|\Omega^1)$ with respect to θ. Consequently, the P_0-expected value of the second term between brackets in the first display is zero, hence the validity of the alternative version of Σ. The conclusion simply follows from another application of Theorem 5.23 in [14] in the classical iid framework associated with $P_1^\star$. □

References

[1] *IARC monographs on the evaluation of the carcinogenic risk of chemicals to man: asbestos*, volume 14. IARC, 1977.

[2] A. Belot, P. Grosclaude, N. Bossard, E. Jougla, E. Benhamou, P. Delafosse, A. V. Guizard, F. Molinié, A. Danzon, S. Bara, A. M. Bouvier, B. Trétarre, F. Binder-Foucard, M. Colonna, L. Daubisse, G. Hédelin, G. Launoy, N. Le Stang, M. Maynadié, A. Monnereau, X. Troussard, J. Faivre, A. Collignon, I. Janoray, P. Arveux, A. Buemi, N. Raverdy, C. Schvartz, M. Bovet, L. Chérié-Challine, J. Estève, L. Remontet, and M. Velten. Cancer incidence and mortality in France over the period 1980-2005. *Rev. Epidemiol. Santé Publique*, 56(3), 2008. Detailed results and comments [online] `http://www.invs.sante.fr/surveillance/cancers/estimations_cancers/default.htm`.

[3] H. K. Biesalski, B. B. de Mesquita, A. Chesson, F. Chytil, R. Grimble, R. J. Hermus, J. Kohrle, R. Lotan, K. Norpoth, U. Pastorino, and D. Thurnham. European consensus statement on lung cancer: risk factors and prevention. Lung cancer panel. *CA Cancer J. Clin.*, 48(3):167–176, 1998.

[4] R. S. Chhikara and Folks J. L. *The inverse Gaussian distribution: theory, methods and applications*. Marcel Dekker: New-York, 1989.

[5] C. A. Haiman, D. O. Stram, L. R. Wilkens, M. C. Pike, L. N. Kolonel, B. E. Henderson, and L. Le Marchand. Ethnic and racial differences in the smoking-related risk of lung cancer. *N. Engl. J. Med.*, 354(4):333–342, 2006.

[6] M.-L. T. Lee and G. A. Whitmore. Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Statist. Sci.*, 21(4):501–513, 2006.

[7] M-L. T. Lee and G. A. Whitmore. Proportional hazards and threshold regression: their theoretical and practical connections. *Lifetime Data Anal.*, 16(2):196–214, 2010.

[8] P. Morfeld. Years of life lost due to exposure: Causal concepts and empirical shortcomings. *Epidemiol. Perspect. Innov.*, 1(1), 2004.

[9] J-C. Pairon, B. Legal-Régis, J. Ameille, J-M. Brechot, B. Lebeau, D. Valeyre, I. Monnet, M. Matrat, and B. Chamming's, S. Housset. Occupational lung cancer: a multicentric case-control study in Paris area. European Respiratory Society, 19th Annual Congress, Vienna, 2009.

[10] J. Robins and S. Greenland. The probability of causation under a stochastic model for individual risk. *Biometrics*, 45(4):1125–1138, 1989.

[11] J. Robins and S. Greenland. Estimability and estimation of expected years of life lost due to a hazardous exposure. *Stat. Med.*, 10(1):79–93, 1991.

[12] S. Rose and M. J. van der Laan. Simple optimal weighting of cases and controls in case-control studies. *Int. J. Biostat.*, 4:Art. 19, 24, 2008.

[13] M. J. van der Laan. Estimation based on case-control designs with known incidence probability. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 2008. Paper 234.

[14] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[15] A. W. van der Vaart, S. Dudoit, and M. J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statist. Decisions*, 24(3):351–371, 2006.

[16] E. A. Zang and Wynder. E. L. Differences in lung cancer risk between men and women: examination of the evidence. *J. Natl. Cancer Inst.*, 88(3-4):183–192, 1996.