Working Paper Series, N. 5, March 2011

Pleural Mesothelioma: forecasts of the death toll in the area of Casale Monferrato, Italy

Claudia Furlan and Cinzia Mortarino
Department of Statistical Sciences
University of Padua
Italy

Abstract: In the city of Casale Monferrato, the largest Italian factory that produced asbestos cement goods was active from 1907 to 1985. As a consequence, asbestos fibers scattered in the surrounding area and caused an enormous number of pleural mesotheliomas. Due to the very long latency of this disease, many subjects have not exhibited its symptoms yet. The aim of this paper is to model and predict the future evolution of the number of deaths due to this disease among residents in the area around that city. The model used here is based on a Cellular Automata that is assumed to pass through three steps: exposure, contamination, diagnosis. In that way, forecasts of the future evolution take into account the environmental conditions that changed in time during the last century because of different levels in plant activity. The model is fitted to annual diagnosis data starting from 1954 to 2009. Results show that deaths will not end until 2033, and that in the next two decades, at least 479 more subjects will be diagnosed with this disease.

Keywords: Asbestos, Cellular Automata, Environmental Exposure, Pleural Mesothelioma, death toll, cancer registries

First Draft (2011-03-30)

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Cinzia Mortarino
tel: +39 049 827 4184
mortarino@stat.unipd.it
http://www.stat.unipd.it/~mortarino

Pleural Mesothelioma: forecasts of the death toll in the area of Casale Monferrato, Italy

Claudia Furlan and Cinzia Mortarino
Department of Statistical Sciences
University of Padua
Italy

Abstract: In the city of Casale Monferrato, the largest Italian factory that produced asbestos cement goods was active from 1907 to 1985. As a consequence, asbestos fibers scattered in the surrounding area and caused an enormous number of pleural mesotheliomas. Due to the very long latency of this disease, many subjects have not exhibited its symptoms yet. The aim of this paper is to model and predict the future evolution of the number of deaths due to this disease among residents in the area around that city. The model used here is based on a Cellular Automata that is assumed to pass through three steps: exposure, contamination, diagnosis. In that way, forecasts of the future evolution take into account the environmental conditions that changed in time during the last century because of different levels in plant activity. The model is fitted to annual diagnosis data starting from 1954 to 2009. Results show that deaths will not end until 2033, and that in the next two decades, at least 479 more subjects will be diagnosed with this disease.

Keywords: Asbestos, Cellular Automata, Environmental Exposure, Pleural Mesothelioma, death toll, cancer registries

1 Introduction

Asbestos is the main causal factor for pleural mesothelioma (PM). One of the areas affected by the epidemic of PM is Casale Monferrato, in the northwest of Italy, where the largest Eternit asbestos cement factory was active from 1907 to 1985. Only in the last decade the effects of asbestos exposure has begun to show its gravity. Indeed, the symptoms of PM appear after a latency whose duration is variable. In Marinaccio *et al.* [18] a mean latency of 45 years is reported, even if lower values are expected for heavier exposures. In Selikoff *et al.* [29] it is reported that observed latency values peak around 45 years with non-decreasing risk for longer latencies, while before 15–19 years the risk is negligible. Hillerdal [14] writes that the latency time varies in different cohorts, and depends on how long a cohort is followed up. Latency time also depends on exposure, varying from 29.6 years for insulators (with the highest exposure) to 51.7 in women with domestic exposure. Because of this variability, a reliable estimate of how long the epidemic will still last and of the expected death toll, still represents a great challenge.

In the following, we briefly review the large body of literature concerning the PM expression. Most of the papers are based on national PM counts which essentially arise as the consequence of professional exposure to asbestos fibers. Conversely, only a smaller group of contributions evaluates PM risk in people exposed for environmental or domestic reasons (i.e., living or working in structures built with asbestos materials).

Among the large number of works dealing with PM, several studies concern individual risk of dying of this disease at different ages and different periods and are traditionally performed with an age-period-cohort (APC) analysis [25] [24] [23] [28] [26] [27] [13] [21]. Data are aggregated at the regional or more often the national level and age-specific death rate are performed for each 5-year calendar period. Predictions of the death toll are then deduced by applying the relative risks to plausible future population scenarios. Mostly, the individual exposure history (such as intensity and duration of exposure) is not included in the model, and only in some works a discussion on asbestos exposure is added. A part of this field deals also, or exclusively, with deaths due to occupational exposure, restricting the analysis to males only, because jobs with asbestos risk were carried out almost exclusively by men. Peto *et al.* [25] rank the different jobs by the proportional mortality ratio. However, the predictions of the risk across the birth cohorts and of the future evolution of the number of death, based on 5-year calendar period, ignore both the effect of different exposure intensities/durations and the latency dynamics. Indeed, the main concerns stem from the reliability of predictions after 2015/2020. Firstly, because they refer essentially to men born at the end of the '50s or later, for whom no data are yet available except for those who developed PM after a very short latency. Secondly, because the increasing rate of the relative risk in the last years includes the effect of a growing awareness and diagnostic ability of PM. With regard to pollution intensity in occupational exposure analysis, most of the works in this field do not take into consideration the asbestos cycle in modeling, ignoring that asbestos was not used throughout the last century consistently with the same intensity and that was banned at the beginning of the '90s from most industrialized countries. At best, different evolutions of asbestos use arising from qualitative considerations are attempted in order to choose the one that fits better to available data. An exception to this is represented by Marinaccio *et al.* [19], who used the Italian national asbestos consumption curve as an indirect measure of exposure in the general population, at the aggregate level, in 1970-1999. Model adequacy measures revealed that 40 years was the optimal lag between the consumption data and the annual number of deaths. This information was then used, in APC analysis, to relate the risk estimates to the predicted evolution of the death trend.

It is worth mentioning an alternative to APC analysis which is proposed by Banaei *et al.* [2]. The authors of this paper studied the mortality from PM in France on a sample of male workers with a known occupational exposure history in 1925-1995. Data on asbestos imports were used to estimate the parameters of a risk function that links a person's risk of dying from PM at a given age with his past exposure to asbestos. However, since their model is developed for male workers only and deals with occupational exposures only, it is not useful for this study, because in our dataset both men and women (not necessarily workers) are included and different

type of exposures coexist. Moreover, the details on individual exposure histories required by [2] are not available.

Papers dealing with environmental exposure are less common. In all of them, the aim is to estimate the incidence rate for people living close to an asbestos mine or to an asbestos goods factory and to decide whether the incidence rate is significantly greater than the average rate in the rest of the population [7] [12] [15]. The method used is often a case-control study. Works of this type in Italy deal with the surroundings of Casale Monferrato. In particular, results in Maule [20] denote that the risk decreased rapidly as the distance from the factory increased, but at 10 km the risk was still 60% of its value at the source. In Magnani *et al.* [16] also there is evidence of a spatial trend in PM risk. Ferrante *et al.* [9] evaluate the risk of PM in a cohort of wives of asbestos workers, proving that it is much greater than the risk for people living in the rest of the region. In Magnani *et al.* [17], incidence analysis is performed after exclusion, from the population, of the workers in the factory and of people known to live with workers. The analysis is performed on residents in the Local Health Authority (LHA) of Casale, and that area is separated into three zones: the city of Casale M., the towns sharing boundaries with it, and the other towns of the LHA. Results of [17] prove that the risk in the whole LHA territory is greater than in the rest of the region, (with similar values for men and women), but because of small number of data, further confirmation is needed about the increasing trend in risk for zones closer to the plant site.

As above explained, traditional studies are focused on measuring a subject's risk (or incidence rates) and deduce the death toll as a consequence. Conversely, in this paper, the aim is to predict *directly* the death toll among residents in the LHA (of both workers and non-workers) and the year when the PM epidemic will end. The challenge of providing reliables estimates is faced by implementing, for the first time in this context, a model born in the framework of epidemic diseases [4], aimed at estimating the date of the end of the epidemic and the total number of "infected" subjects. The model is adapted here to take into account the individual exposure history, that is the duration of the exposure and the pollution level by asbestos in that period. But, differently from APC and case-control analyses, individual details are not necessary because we only need the total number of subjects with the same exposure history, that is the number of subjects that were resident in the same area at the same time. Since no more information is required, it was possible to reenact the time series since 1954 and then forecasts are based on 56 annual data. Moreover, the model is implemented for different zones of the LHA to consider a spatial trend in PM evolution. The probability function of the time from contamination (defined in Sect. 3) to diagnosis is explicitly estimated to give more information about its variability.

The remainder of the paper is structured as follows. Section 2 describes the data sources, Section 3 gives a detailed description of the origin of the model, and Section 4 provides the estimates and the forecasts. Finally, a discussion follows in Section 5.

2 Data sources

In this study, the data consist of annual PM diagnosis counts obtained by integrating three sources. We notice that some sources report the death year, while others report the diagnosis year. The disease has a very poor prognosis, and survival is generally shorter than one year. For this reason, accordingly to other studies, mortality is considered the same as the incidence, and it is fully expected that this approximation does not cause any additional bias. Specifically, the three sources are:

(a) RENAM (NAtional Mesothelioma REgistry), for 1990–2004,

(b) Division of Pathological Anatomy, City Hospital, Casale M. for 1989–2009,

(c) Public Prosecutor's office of Torino with the plaintiffs' list in the proceedings to the managers of Eternit (which owned the plant). The proceedings are presently ongoing, and the managers are charged with omitting prevention tools in order to reduce fiber dispersion, although the danger of cancer onset due to asbestos exposure has been known since the Sixties.

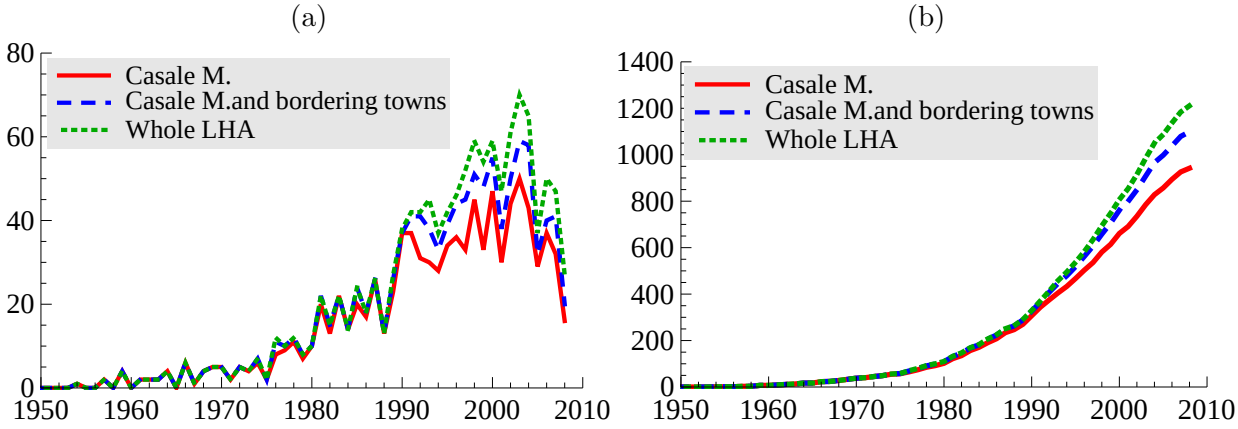
Data arising from source (c) report diagnosis during the whole period of study (the first diagnosis is dated at 1954), but actually the most recent years are not fully covered; the reason is that people who were diagnosed with a PM only recently, did not have the time to complete the legal procedure that is necessary for inclusion as plaintiffs. Indeed, a local newspaper's article published in October 2009, states that the number of uncompleted procedures exceeds 300. Because of this, the diagnosis counts from 2004 until 2009 can be considered to be covered only by source (b). It follows that the annual diagnosis counts of 2004-2009 are underestimated in a non-negligible way and any predictions arising from our study have to be considered only as a lower bound for what we should expect in the next decades.

As mentioned in Sect. 1, the PM risk decreases as the distance from the asbestos source increases. Among the papers dealing with environmental exposure, both [17] and [20] took into account a spatial trend in PM risk. Magnani *et al.* [17] realized the aim by dividing the surroundings of Casale Monferrato in three zones: the city of Casale M., the towns sharing boundaries with it, and the other towns of the Local Health Authority (LHA). Maule *et al.* [20] conversely considered the distance from the plant site. For our study, the address details included in the diagnosis data concern only the city of residence of the subjects at the time of diagnosis, but not a complete address. Because of this, since it is not possible to separate the town territory of Casale M. in several precincts, we follow the division of the LHA of Casale M. adopted by Magnani *et al.* [17], in order to detect how the spatial trend affects diagnosis counts. Fig. 1 shows the annual and the annual cumulative diagnosis counts of PM among residents, respectively, of the city of Casale M., of the city and bordering towns, and of the whole LHA. Evidently, most deaths concern residents in the city of Casale M., while residents in the other towns of the LHA are less affected by this disease, as expected since those towns are farther away from the pollution source.

For this study, we also rely on:

- the number of residents $r(t)$ from 1907 to 1986 in the three zones (source: ISTAT);

Figure 1: Annual (a) and annual cumulative (b) PM diagnosis counts.



- the annual number of asbestos fibers, $A(t)$, processed in the manufacturing plant. The exact values of this time series are by now available only to the Public Prosecutor's office that is responsible for the mentioned proceedings. Provisionally, we use approximate values obtained through processing data about national production, the time-varying amount of that quantity produced from the plant under study, and the average asbestos quantity used for each ton of finished products. When the mentioned legal procedure ends, more precise data will be used, and this source of uncertainty in forecasts will be removed.

3 Model

In this study, data are available only *at the population level,* (i.e., annual numbers of diagnosis), but the connection of PM's expression with time-dependent environmental conditions *at the individual level* is an essential feature in modeling. This is the reason why we choose to ground on a Cellular Automata (CA) model [4]. The most striking feature of such a model is its skill of: defining the *individual* transition rule among the states of the disease; obtaining a *population* model by 'summing up' the individual transition rule; finally, interpreting the parameter estimates again at the *individual* level.

CA models have already been applied in epidemic contexts to model the spread of an infectious disease, both in the case when infected individuals are dead or become permanently immune [5], and in the case when infected individuals, after recovery, become susceptibles again to catch the disease [6]. Our context is different from those described in the latter studies because PM is not an infectious disease and PM implies a latency period that depends on individual characteristics and time-varying environmental pollution. Indeed, the prerequisite for developing PM is to be exposed to asbestos fibers: the bigger ones are eliminated (expectorated or digested), while the smaller ones are transferred to body tissue, where they become permanently lodged. In some individuals, after an unknown period of time, the carcinogenesis

process begins in latent form, and after a long period of time, depending on exposure intensity and duration, symptoms appear and PM is diagnosed.

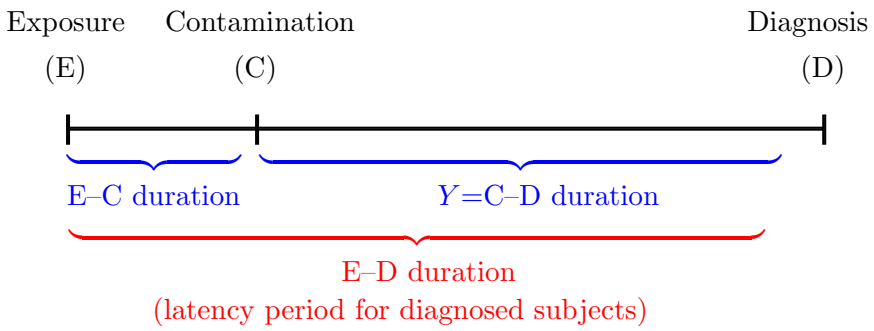
This kind of evolution suggested us to focus on three phases (see Fig. 2): *exposure* (E), which means a subject lived in the LHA for at least one year, *contamination* (C), that is the beginning of the disease in latent form, and *diagnosis* (D), when symptoms appear. It is important to highlight that an exposed subject may never get contaminated, while a contaminated subject will surely develop PM after a long period, whose duration, denoted by Y, varies among individuals (see 'C-D duration' in Fig. 2). For those subjects who will develop PM, the duration of the period between exposure and contamination (denoted by 'E-C duration' in Fig. 2) depends on environmental exposure, while the duration of the whole process from exposure to diagnosis ('E-D duration' in Fig. 2) corresponds to the duration of the latency period.

We underline that the three phases above described should not be confused with the carcinogenesis stages known in the literature ([8], [1]). The latter are a series of sudden and irreversible changes which must take place in a specific order in cells to determine cancer. In our work, we rely on a simplified model of the real process and introduce, for the first time, the term *contamination* to denote the final irreversible time point at which PM takes place in a latent way. In other words, contamination could be interpreted as the final stage in carcinogenesis. That step of course does not correspond to clinical diagnosis because a delay period is required to reach a detectable condition.

The described structure leads to a CA framework with three states (E, C, D), the second of which, differently from [5], [6], is latent. A further difference between our approach and the approach followed in [5], [6], is that we do not resort to simulation techniques, but we validate our model through empirical data.

Guseo & Guidolin [11] used a similar structure (two observable states separated by a latent one) in the context of information diffusion processes. The second latent state was represented by being informed before the adoption (of market goods or

Figure 2: Exposure, Contamination, and Diagnosis phases of PM expression.



services) while adoption was the third observable state. In our work, the latent state is represented by being contaminated, and we focus on modeling the probability of moving from E to C with time-varying environmental conditions. The essential difference is that in [11] the transition from the latent state to the observable state may not happen (informed people may not purchase a good). In our case, the transition is compulsory (a contaminated subject will surely develop PM).

Notice that in this model the factory's workers are not pulled separately from the rest of the population because the aim is not to measure a subject's risk, but to predict the death toll of both workers and non-workers. A different evolution rule for workers and non-workers is one of the matters that will be studied in further developments, for improving the reliability of the forecasts. This idea could not be implemented in the present work because, due to the proceedings to the managers of Eternit, we do not dispose at present of detailed data about the number of workers that were employed in the plant during the production period.

In the following, the contamination and the PM expression processes are illustrated together with the way of moving from the individual to the aggregate level.

3.1 The contamination process

Contamination has been defined, just above, as the beginning of the disease in latent form. An exposed subject may never get contaminated. A contaminated subject will surely develop PM after a period whose duration Y varies from 1 to K years (K unknown). Let us denote with

$$c(i;t) = \begin{cases} 0 & \text{if subject } i \text{ is not contaminated at year } t \\ 1 & \text{if subject } i \text{ is contaminated at year } t \end{cases}$$

the contamination step of a subject. This is a non-observable datum.
Note that, if a subject i is contaminated in year t, then $c(i;s) = 1$ for $s \geq t$ (i.e., contamination is an irreversible transformation). Conversely, for an uncontaminated subject i at year t, the contamination process unfolds according to the following rule:

$$c(i;t+1) \sim Be(p_c(t+1)) \quad \text{if} \quad c(i;t) = 0, \tag{1}$$

where $Be(\cdot)$ denotes a Bernoulli random variable, and $p_c(t+1)$ is the *transition* contamination probability of an individual of being contaminated exactly at year $t+1$, that is the probability of the subject of being contaminated at year $t+1$, given it was not contaminated at year t (or earlier). In other words, it is the probability of moving from state E to state C (Fig. 2) exactly at year $t+1$. This one, for the reasons described above, has to be modelled as a function of the time-varying environmental pollution, that is approximated by the asbestos quantity $A(t)$ used in the plant in year t:

$$p_c(t+1) = \gamma A(t), \tag{2}$$

where γ is a proportionality parameter (that has to be estimated) between the transition contamination probability and the exposure intensity. We underline that $p_c(t+1)$ (probability of *moving* from E to state C exactly at year $t+1$) should not be confused with the state probability of *being* contaminated at year $t+1$, that is

the probability of being in state C at year $t+1$. The latter does not depend only upon exposure intensity but also indirectly upon exposure duration: an individual with longer exposure has a greater state contamination probability. This is the consequence of the fact that the state probability for subject i corresponds to:

$$\begin{aligned} P\left[\text{of being contaminated at } t+1\right] &= P\left[c(i,t+1)=1\right] \\ &= 1 - P\left[\text{not yet being contaminated at } t+1\right] \\ &= 1 - \prod_{s=t_0}^{t} \left[1 - p_c(s+1)\right] \\ &= 1 - \prod_{s=t_0}^{t} \left[1 - \gamma A(s)\right], \end{aligned} \tag{3}$$

where t_0 is the first year of exposure of subject i. The state contamination probability in Eq. (3) represents the risk of developing PM for those individuals that have been living in LHA for a certain number of years, and it takes into account not only the duration of the exposition but also the time-varying intensity exposition. This structure describes the consensus of most researchers agreeing there is a positive dose-response curve for mesothelioma–the stronger the exposure to asbestos, the greater the risk [3] [14]. Moreover, this model is fully coherent with the common belief that longer latencies arise for less intensive exposures [19]. Since the latency is the duration between E and D for diagnosed individuals, a longer latency comes out to be the result of a low contamination probability, which on average leads to a longer 'E-C duration' and consequently to a longer 'E-D duration'.

3.2 PM expression process

As far as the process observable component (PM diagnosis) is concerned, the following notation is used:

$$d(i;t) = \begin{cases} 0 & \text{if subject } i \text{ is not diagnosed with PM at year } t \\ 1 & \text{if subject } i \text{ is diagnosed with PM at year } t. \end{cases}$$

Note that, if a subject i is diagnosed at year t, then $d(i;s)=1$ for $s \geq t$. Conversely, for a subject i undiagnosed at time t, the PM expression's evolution rule is modelled as follows:

$$d(i;t+1) \sim Be(p_{di}(t+1)) \quad \text{if} \quad d(i;t)=0, \tag{4}$$

where $p_{di}(t+1)$ denotes the *transition* probability of a contaminated subject i of being diagnosed at year $t+1$, that is the probability of the subject being diagnosed at year $t+1$, given that he was not diagnosed at year t (or earlier). In other words, it is the probability of moving from state C to state D exactly at year $t+1$.

The p_{di} probability depends upon i because it has to take into account the contamination process of subject i. In order to connect the C state to the D state, let us denote by Y the random variable measuring the duration of the period from contamination to diagnosis ('C-D duration' in Fig. 2). We assume that Y can take values in $\{1,2,\ldots,K\}$ where K is a parameter that represents the maximum feasible

length of C-D period. We choose a Binomial distribution for the C-D duration, that is $Y \sim 1 + Bin(K-1, p)$, whose probability function is:

$$P[Y = y] = P[Bin(K-1, p) = y-1] = \binom{K-1}{y-1} p^{y-1}(1-p)^{K-y}. \tag{5}$$

For any subject i, contaminated from y years (at year $t-y+1$), the probability of a diagnosis at year $t+1$ is:

$$p_{di}(t+1) = P[d(i;t+1) = 1 \mid d(i;t) = 0,\ c(i;t-y+1) = 1,\ c(i;t-y) = 0]$$

and this expression equals the hazard rate of Y evaluated in y, since it corresponds to:

$$p_{di}(t+1) = P[Y = y \mid Y \geq y] = \frac{P[Y = y]}{P[Y = y] + P[Y = y+1] + \ldots + P[Y = K]}.$$

A sensible assumption for $p_{di}(t+1)$ is to make it depending on the *number of years* already passed from the year of contamination:

$$p_{di}(t+1) = \sum_{k=0}^{K-1} c(i;t-k)\mu_{k+1}, \tag{6}$$

where μ_{k+1} are a set of parameters. In this way, if subject i is contaminated since y years, $p_{di}(t+1)$ would be equal to $\mu_1 + \mu_2 + \ldots + \mu_y$, for $y \leq K$, because

$$c(i;t) = \cdots = c(i;t-y+1) = 1 \quad \text{and} \quad c(i;t-y) = \cdots = c(i;t-K+1) = 0.$$

In the case of contamination since K years (maximum feasible of C-D duration), the hazard rate is 1 by definition, and then:

$$p_{di}(t+1) = \sum_{k=0}^{K-1} \mu_{k+1} = 1.$$

Conversely, if subject i is not yet contaminated at year t, that is $c(i;t) = 0$ and, consequently, $c(i;t-k) = 0 \ \ \forall k$, he cannot be diagnosed for sure at year $t+1$ and $p_{di}(t+1)$ results to be 0.

With the chosen parametrization in Eq. (6), μ_{k+1} represents the increase in risk of being diagnosed in the transition from k to $k+1$ years from contamination. Since the Binomial distribution is an IHR (increasing hazard rate) distribution, ([22], [10]), assumption (5) entails that μ_k values are nonnegative.

3.3 From the individual to the aggregate level

Both the contamination and the PM expression processes are defined by individual transition rules (through Eqs. (1) and (4)) to move, respectively, from exposure to contamination and from contamination to diagnosis. Modeling the rules at the individual level comes out to be necessary for this phenomenon, since the time of transition from state C to state D depends on the individual contamination state

which, in turn, depends on individual exposure histories (that is, duration of the exposure and pollution level by asbestos in that period). However, as already mentioned in Section 2, data at the individual level (for *each* subject that ever lived in the LHA) are not available. A useful method, both to preserve the individual rules and using the available annual counts data, is to aggregate the individual rules themselves at the population level by summing up with respect to all subjects that at a certain year were part of the population. In this way, the 'mean' behavior of the whole population is adequately taken into consideration.

Let us consider the aggregation of the contamination process which was described in Subsect. 3.1. Let $C(t)$ be the sum with respect to i of Eq. (1):

$$C(t) = \sum_i c(i;t),$$

that is, for each year t, the cumulative number of contaminated subjects; with assumption of Eq. (2), and using expectation, we obtain that:

$$C(t+1) \simeq C(t) + \gamma A(t) Pop_{risk}(t), \tag{7}$$

where $Pop_{risk}(t)$ is the population at risk for contamination at a given year t. $Pop_{risk}(t)$ is made up, for each year t, by all those individuals that were residents in LHA in a specific year t, except those already contaminated. However, not all the contaminated subjects $C(t)$ should be used to define $Pop_{risk}(t)$, but only the subset constituted by those contaminated still resident in LHA in that year t. We reasonably assume that the rate of contaminated residents in a specific year t, intended as the number of contaminated subjects on the number of residents, corresponds to the rate for cumulative values. This hypothesis leads to the definition of the following population at risk:

$$Pop_{risk}(t) = r(t) - r(t)\frac{C(t)}{R(t-1)} = r(t)\left[1 - C(t)/R(t-1)\right], \tag{8}$$

where $r(t)$ denotes the number of residents in the area in year t, $R(t-1)$ denotes the cumulative number of residents in the area until time $t-1$, and $C(t)/R(t-1)$ defines the rate of contaminated residents at year t. Substitution of $Pop_{risk}(t)$ of Eq. (8) in Eq. (7) leads to a recurrence definition of the process that models the contamination step:

$$C(t+1) \simeq C(t)\left[1 - \gamma\frac{A(t)r(t)}{R(t-1)}\right] + \gamma A(t) r(t), \tag{9}$$

with initial value $C(1) = \gamma A(0) r(0)$.

Let us consider now the aggregation of the process that describes the disease's expression of Subsection 3.2. If we denote the sum with respect to i of Eq. (4) by

$$D(t) = \sum_i d(i;t),$$

that is, the cumulative number of diagnosed subjects at year t, it leads to

$$D(t+1) = D(t) + \sum_i Be(p_{di}(t+1))\, I_{\{d(i;t)=0\}}, \tag{10}$$

where $p_{di}(t+1)$ is defined in Eq. (6) and depends on K, $\{\mu_{k+1}\}$, and γ (through the contamination process, $c(i;t)$), while I_A is the indicator function of the event A.

In the Appendix, it is proven that previous expression (10) results to be:

$$D(t) \simeq \sum_{y=1}^{\min(K,t-1)} P\left[Y=y\right]\ C(t-y) \qquad \forall t \geq 2, \tag{11}$$

where Y is distributed as described in Eq. (5), and $C(t-y)$ is defined in Eq. (9), with $D(0) = D(1) = 0$. Note that, $D(t+1)$ in Eq. (10) depends on parameters $(K,\ \{\mu_k\})$, while, through (5), it depends on parameters $(K,\ \gamma,\ p)$; besides, it is a function of the known time series $A(t)$, $r(t)$ and $R(t-1)$.

The apparent complexity of the model structure is overcome by the simply and sensible interpretation for the final expression of Eq. (11). The C-D duration, Y, for contaminated subjects varies with probabilities $P\left[Y=y\right]$ of Eq. (5), for $y = 1,\ldots,K$. That is, subjects contaminated exactly at year $t-1$ will be diagnosed at year t with probability $P\left[Y=1\right]$, those contaminated exactly at year $t-2$ will be diagnosed at year t with probability $P\left[Y=2\right]$, etc.. So, we obtain that:

$$\begin{aligned} D(t) - D(t-1) = & \\ \overset{(11)}{=} & \sum_{y=1}^{\min(K,t-1)} P\left[Y=y\right]\ C(t-y) - \sum_{y=1}^{\min(K,t-2)} P\left[Y=y\right]\ C(t-1-y) \\ = & \sum_{y=1}^{K} P\left[Y=y\right]\ \left[C(t-y) - C(t-1-y)\right] \quad \text{for } \ t \geq K+2, \end{aligned}$$

i.e., the number of *new* diagnoses realized at year t, is represented by those subjects contaminated exactly at year $t-1$, or $t-2$, and so on, with a C-D duration of respectively 1, 2, ..., years. In a dual way, the group of diagnosed subjects at time t, is composed by a mixture of subjects contaminated at different time points, and the "mixing rule" is given by the probabilities of Y distribution.

If we denote by $w(t)$ the observed values of the time series counting the cumulative annual number of diagnosis, we end up fitting the following model:

$$w(t) = D(t) + \varepsilon. \tag{12}$$

Due to the complex structure of function $D(\cdot)$, in order to concretely evaluate parameter estimates of this regression model, K was in turn kept fixed, and estimates for γ and p were evaluated conditionally based on the choice of K. In the end, the optimal K value was chosen as the one that made the residual deviance smaller:

$$\hat{K} \ = \ \arg\min_K SSE(K) = \arg\min_K \left\{ \min_{\gamma,p} \sum_t \left[w(t) - D(t)\right]^2 \right\}. \tag{13}$$

Figure 3: Probability function for the duration Y of C-D period and transition contamination probability $p_c(t+1)$.

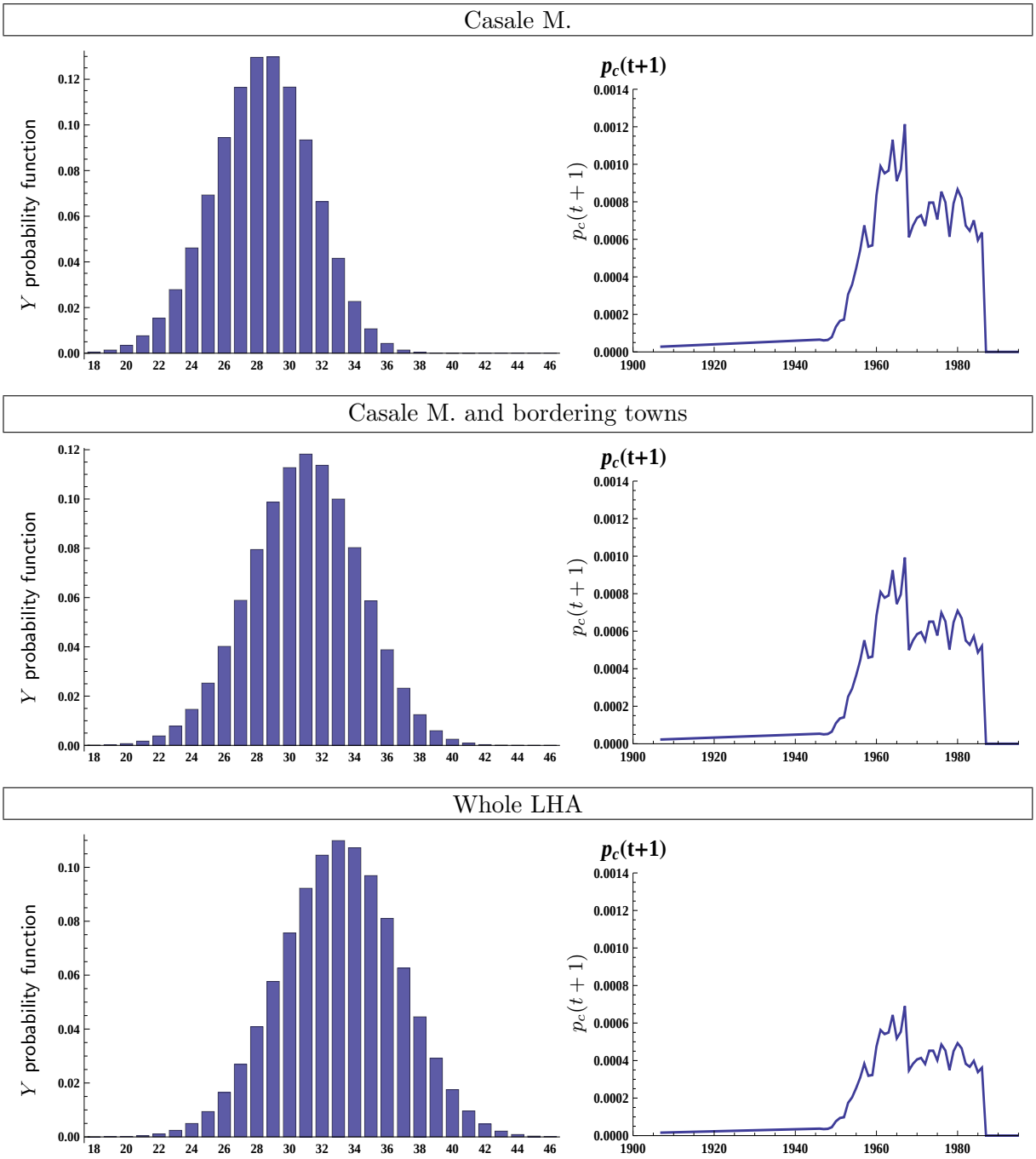


Table 1: Estimates, standard errors and 95% CI of parameter estimates. R^2 for goodness-of-fit.

		Estimate	Std. Error	95% CI
Casale M.	γ	1.8598×10^{-8}	2.5756×10^{-10}	$(1.8087, 1.9109)\times10^{-8}$
($\hat{K}=42,\ R^2=0.9959$)	p	0.6670	0.0059	(0.6554, 0.6787)
Casale and bordering towns	γ	1.5361×10^{-8}	2.6810×10^{-10}	$(1.4829, 1.5892)\times10^{-8}$
($\hat{K}=49,\ R^2=0.9943$)	p	0.6235	0.0061	(0.6114, 0.6356)
Whole LHA	γ	1.0575×10^{-8}	2.4140×10^{-10}	$(1.0096, 1.1054)\times10^{-8}$
($\hat{K}=55,\ R^2=0.9916$)	p	0.5941	0.0068	(0.5805, 0.6076)

Table 2: Forecasts about the C-D duration (average), the year when the diagnosis will end, the number of future diagnoses, the total number of diagnoses. Observed number of diagnoses until 2009 is also given.

	Casale M.	Casale M. and bordering towns	Whole LHA
Average C-D duration	28.35	30.93	33.08
Year of PM end	2025	2028	2033
Obs. diagnoses until 2009	942	1099	1211
Future diagnoses	208	353	479
Total diagnoses	1150	1452	1690

4 Results

As anticipated in Sect. 2, the analysis was performed separately for each of the three zones there presented. First, the optimal value of K was chosen by Eq. (13), and then model (12) was fitted to the time series of annual diagnosis counts. We remind that K denotes the largest admissible value for Y. It assumes value $\hat{K}=42$ for Casale M., $\hat{K}=49$ for Casale M. and bordering towns, and $\hat{K}=55$ for the whole LHA. Table 1 also shows a trend in $\hat{\gamma}$ and $\hat{p}$, which both decrease as the limit of the territory is expanded. This feature means that the transition contamination probability $p_c(t+1)$ for an exposed subject is reduced, while the average and the maximum value for the C-D duration (and then for the latency) are longer.

The estimated probability function for C-D duration and the contamination probability are displayed in Fig. 3. The picture gives evidence of an annual contamination probability that has been greater than 0.0006 in Casale for more than 30 years (from 1955 to 1985) with a peak in the mid-sixties of 0.012. In the same period, the contamination probability has been greater than 0.0005 in Casale M. and bordering towns and greater than 0.00035 in the whole LHA, with a peak of 0.0010 and 0.0007 respectively. Through the state probability of Eq. (3), it is now possible to quantify the risk of a subject of developing PM, based on his individual exposure history. For instance, for a subject that lived in Casale M. from 1960 to 1986, the probability of being contaminated (and then of surely develop PM) is 0.0214.

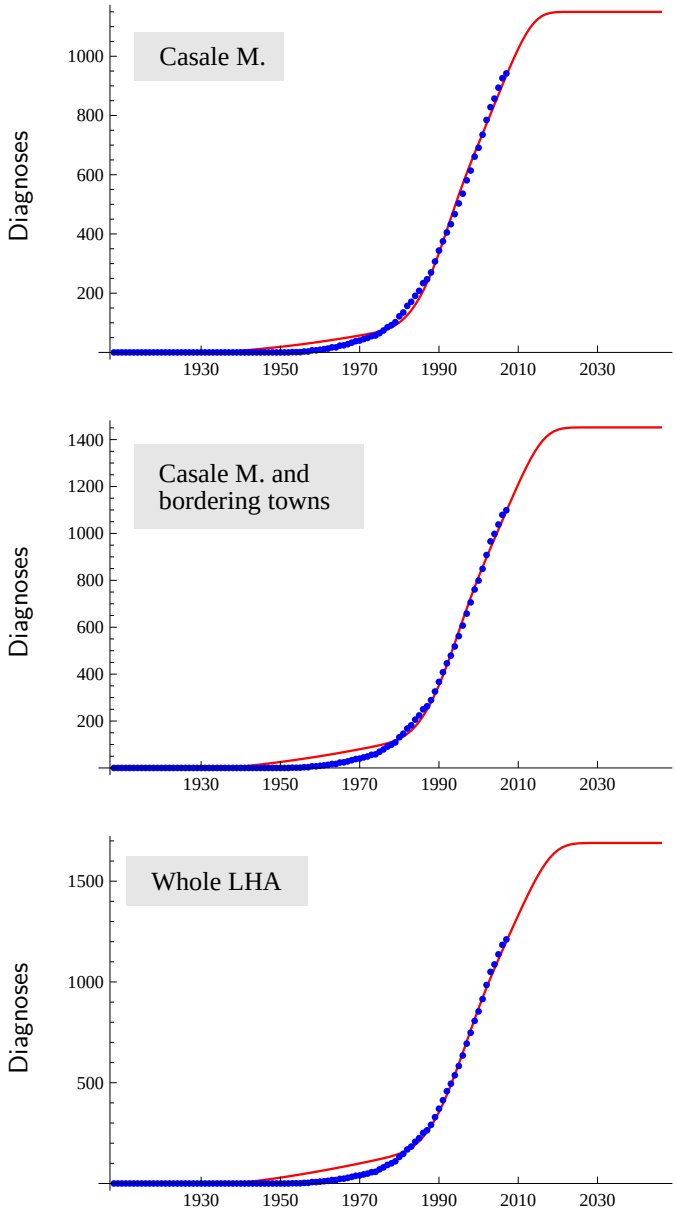
The estimated probability function of Y, in Fig. 3, shows non-negligible proba-

bilities associated with C-D duration between 18 and 37 years in Casale M., with a peak round 28/29 years and the average length equals to 28.35 years (see also Table 2). With respect to Casale M. and bordering towns and the whole LHA, non-negligible probabilities are associated between 20 and 41 years and between 21 and 45, respectively, with a peak round 31 and 33 years and the average length that increases from 30.93 to 33.08 years respectively. As a consequence of the previous statements on the spatial changes of the contamination probability and on the C-D duration, the year forecasted as the end of this 'epidemic' moves from 2025 for Casale M., to 2028 to Casale M. and bordering towns, and to 2033 to the whole LHA (Table 2).

Table 2 also displays the number of observed diagnoses until 2009, of future diagnoses and the total number of diagnoses (evaluated as the sum of the observed and of the future diagnoses). In the next years, 208 new diagnoses are expected among people that in the past were residents in Casale M., and according to this model, starting from 2025, the PM epidemic will run out. The figures of the table imply that the bordering towns of Casale M. counted 157 diagnoses in 2009 and 145 new diagnoses are expected before 2028, for a total number of 302 diagnoses. Analogously, for the territory of the LHA except Casale M. and the bordering towns, the observed count of diagnoses was 112 until 2009, and 126 future diagnoses are expected before 2033, for a total number of 238 diagnoses. We remind that these values should be considered as a lower bound of predictions, since the data counts of the last few years are covered by only one out of three sources of data (see Section 2). It should be noticed that while in Casale M. the 82% of total diagnoses was already reached in 2009, only the 52% was achieved in the bordering towns of Casale M. and only the 47% in the territory of LHA except Casale M. and bordering towns. This is a consequence of the fact, already presented in Section 3.1, that lower contamination probability leads to longer E-C duration (and then latency).

The adequacy of model fitting can be evaluated both by the large values of R^2 in Table 1 and by comparison of observed and fitted values in Fig. 4. The adequacy is very good in the second part of the time series, while some discrepances can be observed from 1947 to 1972. In particular, the number of deaths predicted by our model is much greater than reported by the sources. This result can be explained by noticing that PM at that time was poorly known. Moreover, many workers who will have developed PM, died earlier because of the debilitating effects of asbestosis (and the recorded official cause of death was heart failure); furthermore, PM is underestimated in elderly people because they rarely submit to the invasive diagnostic procedures necessary to have an uncontroversial diagnosis. For this reason, in some works (see, e.g., [2] and [19]) the observed number of diagnosis is corrected through some multiplicative factors that substantially take into account that in the past PM diagnosis were underestimated. In this work, we did not take into consideration these corrections because these methods are often subjective and do not always agree. We postpone the inclusion of a censoring effect in the model to future steps of this research.

Figure 4: Model fit for the three zones. Points display the observed cumulative diagnoses $w(t)$, and lines the fitted $D(t)$.



5 Discussion

In this paper, we attempted to provide reliable estimates of the death toll and of the future evolution of the PM, taking into account the long-term dynamics of the latency due to time-dependent exposure to asbestos. The analysis was performed for the residents in the city of Casale M., for the residents in Casale M. and in the towns sharing boundaries with it, and in the whole LHA. The time series of annual diagnosis counts are available from 1954 to 2009, by integrating three sources: RENAM, the Division of Pathological Anatomy of the City Hospital in Casale M., and the plaintiffs' list in the proceedings to the managers of Eternit of the Public Prosecutor's office of Torino. In spite of this, the last period from 2004 until 2009 is covered only by the Division of Pathological Anatomy, and, as a consequence, the predictions furnished by this study have to be considered only as a lower bound for what we should expect in the next decades.

The model used here is a CA with data at the population level, but the connection of the PM expression with time-dependent environmental condition is specified at the individual level. The steps in the evolution of PM expression are: distinct in exposure, contamination, and diagnosis. Of these three phases, only the contamination is latent. An exposed subject is contaminated at year $t+1$ with a probability that is given by the transition rule $p_c(t+1)$, which depends on the asbestos quantity $A(t)$ used in the plant in year t. While an exposed subject may never get contaminated, a contaminated subject will surely develop the PM after Y years. The distribution of the duration Y is specified by a Binomial variable. The transition rule $p_{di}(t+1)$ of a contaminated subject of being diagnosed at year $t+1$ corresponds to the hazard rate of Y. The duration of the whole process from exposure to diagnosis is equivalent to the duration of the latency period.

The results arising from fitting our model to the three territories suggest that, as far as residents in more distant towns from the plant are included, the estimated value for γ is lower, that is the contamination probability for an exposed subject is reduced; the average and the maximum value for the C-D duration are longer; and the year forecasted as the end of this epidemic is moved far away. A larger number of not-yet-diagnosed cases are still expected: 208 new diagnoses in Casale M. before 2025, 145 in the bordering towns of Casale M. before 2028, and 126 in the remaining territories of the LHA before 2033. The fact that lower contamination probability leads to longer E-C duration, and then longer latency, is confirmed by the decreasing percentage of cumulative diagnoses in 2009 on the total number of forecasted diagnoses, as the distance increases from the plant site: in 2009, the observed cumulative diagnoses represent the 82% of total expected diagnoses in Casale M. , while only the 52% in the bordering towns of Casale M., and the 47% in the remaining territories of the LHA.

As a final remark, we underline that the distribution of the C-D duration, that was here specified as a Binomial one, can be changed with a different distribution if required. Moreover, future steps of this research will concern a censoring effect. That effect could be added to the model, approximately to 1947-1972, to solve in that period the apparent overestimation of the number of diagnoses, due to a poor knowledge at that time of the PM.

Appendix

Proof of Equation (11).
If we take the expectation of (10) and substitute (6) into it, through some passages, it can be proven that, for $t \geq 2$,

$$D(t+1) \simeq D(t) + \sum_{k=0}^{K-1} \mu_{k+1} \left[C(t-k) P(Y \geq k+1) - \sum_{j=k}^{K-1} C(t-j-1) P[Y = j+1] \right],$$

while $D(0) = D(1) = 0$.

Starting from $P[Y = 1] = \mu_1$, and

$$\frac{P[Y = y]}{1 - \sum_{k=1}^{y-1} P[Y = k]} = \mu_1 + \mu_2 + \cdots + \mu_y, \quad y = 2, 3, \ldots, K,$$

with simple passages, the relationship linking $P[Y = y]$ values to μ_k values can be proven that for $y = 2, 3, \ldots, K$

$$P[Y = 1] = \mu_1 \qquad \text{and} \qquad P[Y = y] = \left(\sum_{j=1}^{y} \mu_j \right) \left[\prod_{r=1}^{y-1} \left(1 - \sum_{h=1}^{r} \mu_h \right) \right]. \tag{14}$$

By using Eq. (14), the final expression (11) for $D(t)$ can be deduced.

References

[1] P. Armitage. Multistage models of carcinogenesis. *Environmental Health Perspectives*, 63:195–201, 1985.

[2] A. Banaei, B. Auvert, M. Goldberg, A. Gueguen, D. Luce, and S. Goldberg. Future trends in mortality of French men from mesothelioma. *Occup Environ Med*, 27:488–494, 2000.

[3] G. Berry. Models for mesothelioma incidence following exposure to fibers in terms of timing and duration of exposure and the biopersistence of the fibers. *Inhal Toxicol*, 11:111–130, 1999.

[4] N. Boccara. *Modeling Complex Systems*. Springer–Verlag, New York, 2004.

[5] N. Boccara and K. Cheong. Automata network SIR models for the spread of infectious diseases in populations of moving individuals. *J Phys A: Math Gen*, 25(9):2447–2461, 1992.

[6] N. Boccara and K. Cheong. Critical behaviour of a probabilistic automata network SIS model for the spread of an infectious disease in a population of moving individuals. *J Phys A: Math Gen*, 26(15):3707–3717, 1993.

[7] V. Bourdès, P. Boffetta, and P. Pisani. Environmental exposure to asbestos and risk of pleural mesothelioma: review and meta-analysis. *Eur J of Epidemiol*, 16:411–417, 2000.

[8] R. Doll and R. Peto. *The causes of cancer. Appendix E.* Oxford University Press, Oxford, 1981.

[9] D. Ferrante, M. Bertolotti, A. Todesco, D. Mirabelli, B. Terracini, and C. Magnani. Cancer mortality and incidence of mesothelioma in a cohort of wives of asbestos workers in Casale Monferrato, Italy. *Environ Health Perspect*, 115(10):1401–1405, 2007.

[10] P. L. Gupta, R. C. Gupta, and R. C. Tripathi. On the monotonic properties of discrete failure rates. *J Stat Plann and Inf*, 65:255–268, 1997.

[11] R. Guseo and M. Guidolin. Modelling a dynamic market potential: A class of automata networks for diffusion of innovations. *Technol. Forecast. and Soc. Change*, 76(6):806–820, 2009.

[12] J. Hansen, N.H. de Klerk, A.W. Musk, and M.S.T. Hobbs. Environmental exposure to crocidolite and mesothelioma. *Am J Respir Crit Care Med*, 157:69–75, 1998.

[13] K. Hemminki and X. Li. Mesothelioma incidence seems to have leveled off in Sweden. *Int J Cancer*, 103:145–146, 2003.

[14] G. Hillerdal. Mesothelioma: cases associated with non-occupational and low dose exposures. *Occup. Environ. Med.*, 56:505–513, 1999.

[15] D. Luce, I. Bugel, P. Goldberg, M. Goldberg, C. Salomon, M. Billon-Galland, J. Nicolau, P. Quénel, J. Fevotte, and P. Brochard. Environmental exposure to tremolite and respiratory cancer in New Caledonia: A case-control study. *Am J Epidemiol*, 151(3):259–265, 2000.

[16] C. Magnani, P. Dalmasso, A. Biggeri, C. Ivaldi, D. Mirabelli, and B. Terracini. Increased risk of malignant mesothelioma of the pleura after residential or domestic exposure to asbestos: A case–control study in Casale Monferrato, Italy. *Environ Health Perspect*, 109(9):915–919, 2001.

[17] C. Magnani, B. Terracini, C. Ivaldi, M. Botta, A. Mancini, and A. Andrion. Pleural malignant mesothelioma and non-occupational exposure to asbestos in Casale Monferrato, Italy. *Occup. Environ. Med.*, 52:362–367, 1995.

[18] A. Marinaccio, A. Binazzi, G. Cauzillo, D. Cavone, R. De Zotti, P. Ferrante, V. Gennaro, G. Gorini, M. Menegozzo, C. Mensi, E. Merler, D. Mirabelli, F. Montanaro, M. Musti, F. Pannelli, A. Romanelli, A. Scarselli, R. Tumino, and Italian Mesothelioma Register (ReNaM) Working Group. Analysis of latency time and its determinants in asbestos related malignant mesothelioma cases of the italian register. *Eur J of Cancer*, 43:2722–2728, 2007.

[19] A. Marinaccio, F. Montanaro, M. Mastrantonio, P. Uccelli, R. Altavista, M. Nesti, A. Seniori Costantini, and G. Gorini. Predictions of mortality from pleural mesothelioma in Italy: a model based on asbestos consumption figures supports results from age-period-cohort models. *Int. J. Cancer*, 115:142–147, 2005.

[20] M. Maule, C. Magnani, P. Dalmasso, D. Mirabelli, F. Merletti, and A. Biggeri. Modeling mesothelioma risk associated with environmental asbestos exposure. *Environ. Health Perspect.*, 115(7):1066–1071, 2007.

[21] E. Merler, C. Lagazio, and A. Biggeri. Andamento temporale della mortalità per tumore primitivo pleurico e incidenza del mesotelioma pleurico in Italia: una situazione particolarmente grave. *Epidemiol. Prev.*, 23:316–326, 1999.

[22] J. K. Patel. Hazard rate and other classifications of distributions. In S. Kotz and N. L. Johnson, editors, *Encycl of Stat Sci*. Wiley, New York, 2006.

[23] C. Pelucchi, M. Malvezzi, C. La Vecchia, F. Levi, A. Decarli, and E. Negri. The mesothelioma epidemic in western Europe: an update. *British Journal of Cancer*, 90:1022–1024, 2004.

[24] J. Peto, A. Decarli, C. La Vecchia, and E. Negri. The European mesothelioma epidemic. *British Journal of Cancer*, 79:666–672, 1999.

[25] J. Peto, J. T. Hodgson, F. E. Matthews, and J. R Jones. Continuing increase in mesothelioma mortality in Britain. *The Lancet*, 345:535–539, 1995.

[26] B. Price. Analysis of current trends in United States mesothelioma incidence. *Am J Epidemiol*, 145(3):211–218, 1997.

[27] B. Price and A. Ware. Mesothelioma trends in the United States: An update based on surveillance, epidemiology, and end results program data for 1973 through 2003. *Am J Epidemiol*, 159(2):107–112, 2004.

[28] O. Segura, A. Burdorf, and C. Looman. Update of predictions of mortality from pleural mesothelioma in the Netherlands. *Occup Environ Med*, 60(1):50–55, 2003.

[29] I. J. Selikoff, E. C. Hammond, and H. Seidman. Latency of asbestos disease among insulation workers in the United States and Canada. *Cancer*, 46:2736–2740, 1980.

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it
Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Department of Statistical Sciences
University of Padua
Italy