Sensitivity analysis on a Cellular Automata model for the diffusion of Pleural Mesothelioma

Claudia Furlan and Cinzia Mortarino

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 cases of pleural mesothelioma (PM). The model used here, fitted to cumulative annual diagnosis data from 1954 to 2008, takes into account the environmental conditions that changed over the last decades. The modeling approach is in principle appropriate for this type of study. However, it has some limitations that could have an impact on the uncertainty of the forecasts from the model. Mainly, the issues refer to the quality of the data, due to the fact that two of the three available diagnosis sources had not been updated after 2004, and to the use of a proxy for the annual number of fibers, since the true values are not available. We will study how sensitive the overall predictions are to the assumptions made, and the effects of these choices on forecasts. Prediction confidence bands are discussed.

Key words: sensitivity analysis; environmental exposure; pleural mesothelioma; death toll; cancer registries

1 Introduction

The aim of this paper is to discuss the adequacy of the PM death toll's predictions obtained in [1],[2]. The prerequisite for developing PM is to be exposed to asbestos fibers. 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. In this study, data are available only *at the population level* (i.e., annual numbers of diagnoses), but the connection of PM 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 [3]. The data consist of annual PM diagnosis counts, obtained by integrating the three following sources which cover the Local Health Authority (LHA):

Claudia Furlan and Cinzia Mortarino
Department of Statistical Sciences, Padua e-mail: furlan,mortarino@stat.unipd.it

(a) RENAM (NAtional Mesothelioma REgistry), for 1990–2004,
(b) Division of Pathological Anatomy, City Hospital, Casale M., for 1989–2008,
(c) Public Prosecutor's office of Torino with the plaintiffs' list in the proceedings to the managers of Eternit (which owned the plant).

The model describes three steps: exposure (E), contamination (C), and diagnosis (D). In the following, we display only the final expression of the model, while for the model construction we refer to [1],[2]. We denote by $D(t)$ and $C(t)$ the cumulative number of, respectively, diagnosed and contaminated subjects at year t, and by $Y \sim 1 + Bin(K-1, p)$ the distribution of the C–D duration in years. If we denote by $w(t)$ the observed values of the time series counting the cumulative annual number of diagnoses, and by K the maximum feasible length of the C–D period, we fit the following model:

$$w(t) = D(t) + \varepsilon \quad \text{where} \quad D(t) \simeq \sum_{y=1}^{\min(K,t-1)} P[Y = y]\, C(t-y) \qquad \forall t \geq 2. \quad (1)$$

$C(\cdot)$, which is not observable, is defined as: $C(t+1) \simeq C(t) + \gamma A(t) Pop_{risk}(t)$, where $A(t)$ is a proxy for the annual number of fibers, and $Pop_{risk}(t)$ is the population at risk for contamination in a given year t. $Pop_{risk}(t)$ is made up, for each year t, by all those individuals that were residents in the LHA in a specific year t, except those already contaminated [2]. Note that $D(t)$ depends on parameters $(K,\ \gamma,\ p)$.

2 Sensitivity analysis

The modeling approach is in principle appropriate for this type of study. However, it has some limitations that could have an impact on the uncertainty of the forecasts from the model. The first issue concerns the missing values of the count data, in the last four years, due to the fact that sources (a) and (c) had not been updated after 2004, while the second one the use of a proxy for the annual number of fibers (the true values are not available). Both issues are associated with the commencement of legal proceedings by the Public Prosecutor's office against the managers of the factory.

2.1 Inflated data

We have already outlined that the dataset is essentially covered only by source (b) for years 2005-2008. This obviously leads to underestimating the predictions and, in particular, the PM death toll in the three zones. To evaluate the effects of the missing data, the percentage of coverage of source (b) on the count data was evaluated in the five years before (2000-2004). Since the source covered 50.33% of diagnoses in the previous five years, we inflated the data from 2005 to 2008 by dividing by the same percentage. For each zone, Figure 1 shows the observed (lower line) and the inflated data (upper line) that separate after 2004.

Model (1) was then applied to the inflated data (see Table 1 and Figure 2 for the results). The differences of results, in terms of death toll, is that the model with inflated data predicts 268 more diagnoses in the entire LHA (1984 vs. 1716), of

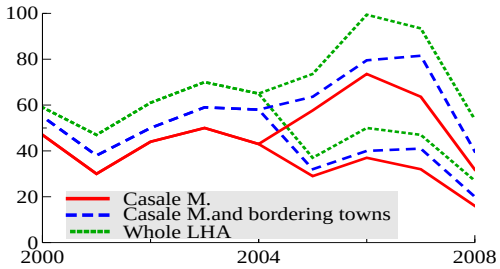

Fig. 1 Observed (lower line) and inflated (upper line) annual count data in the 2000-2008.

which 133 are in the territory of Casale M. With the inflated data, the percentage of total diagnoses already reached in the territory remains broadly constant for Casale M. (approximately 80%), while it slightly decreases for both the bordering towns of Casale M. (which move from 52% to 49%) and the territory of the LHA excluding Casale M. and bordering towns (which moves from 47% to 44%). The expected year PM diagnoses will end, with the model with inflated data, remains 2027 for Casale M., while it is postponed from 2028 to 2032 for Casale M. and bordering towns, and from 2031 to 2034 for the entire LHA.

In order to measure uncertainty associated with the forecasts from the model, confidence bands for each zone were built as follows. The nonlinear estimation procedure provides the joint confidence ellipsoid for the parameters (γ, p), which obviously differs from the rectangular region arising from the combination of the marginal confidence intervals of (γ, p), because of the correlation between them. A grid of about 600 points uniformly covering the 95% confidence ellipsoid was used to evaluate predictions about the death toll. The predictive profiles are plotted, for Casale M. only, in Figure 2 (right picture) for the 2000-2040 time window and can be used as predictive confidence bands. Estimation is quite precise, since the bands are rather narrow.

Table 1 Observed and Inflated data. Forecasts of average C–D duration, year when diagnoses will end, number of future diagnoses, and total number of diagnoses.

	Observed			*Inflated*		
	Casale M.	Casale M. and bordering towns	Entire LHA	Casale M.	Casale M. and bordering towns	Entire LHA
Average C–D duration	28.81	31.27	33.36	30.20	32.78	35.15
Year PM diagnoses will end	2027	2028	2031	2027	2032	2034
Obs. diagnoses up to 2008	942	1099	1211	1055	1230	1370
Future diagnoses	236	379	505	256	435	614
Total diagnoses	1178	1478	1716	1311	1665	1984

2.2 Approximation of the annual number of asbestos fibers $A(t)$

The approximation of the annual number of asbestos fibers is a source of uncertainty for the model. To evaluate how $A(t)$ could influence the predictions, we explored three reasonable scenarios, allowing $A(t)$ to change by a percentage of error ranging

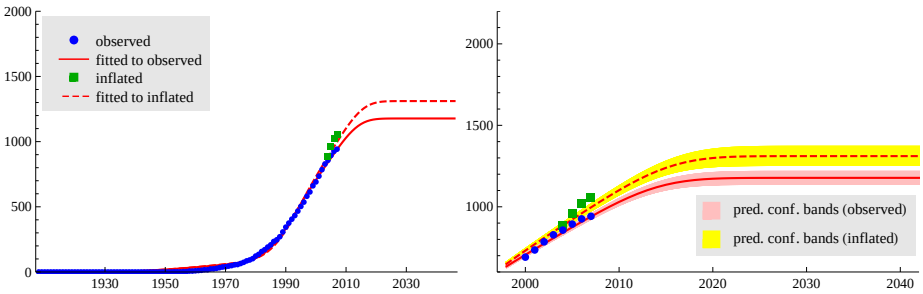

Fig. 2 Casale M.: points display the observed cumulative diagnoses, $w(t)$, while squares display the inflated data. Lines represent the fitted $D(t)$ in both cases. Right picture show the 95% interval bands for the predictions, in the 2000-2040.

Table 2 P-values of the Wilcoxon Test, performed for $e = 10\%,\ 15\%,\ 20\%$ in the three zones.

	10%	15%	20%
Casale M.	0.557	0.322	0.232
Casale and bordering towns	0.625	0.625	0.164
Entire LHA	0.770	0.846	0.160

from 10% to 20%. For each error size, e, we simulated 10 values from a Uniform distribution $U[\cdot,\cdot]$ as follows:

$$A^*(t) \sim A(t) + U[-eA(t), +eA(t)], \qquad e = 0.1,\ 0.15,\ 0.20. \tag{2}$$

For each error size and for each simulation, we then applied the model to the observed data, using $A^*(t)$ instead of $A(t)$. At this point, we compared the predicted death tolls, obtained for each error size, with the corresponding observed death tolls of Table 1, through the Wilcoxon Test. Table 2 shows that there are no significant consequences on the predictions of the death tolls, even if the p-values are smaller for the biggest error size, as expected. In addition, the variability of C–D durations increases slightly with the error size, but, at worst, the difference is only six months. These outcomes point out that our results are quite robust with respect to possible mistakes in the assessment of asbestos fibers.

References

1. Mortarino, C.: Model for diffusion of innovation and Cellular Automata: an epidemiological application to pleural mesothelioma, Proc. 45th SIS Scient. Meeting. Padua, June 16-18 (2010)
2. Furlan, C. & Mortarino, C.: Pleural Mesothelioma: forecasts of the death toll in the area of Casale Monferrato, Italy. WP Series n.5. Dept. of Statistical Sciences, Univ. of Padua (2011)
3. Boccara N. *Modeling Complex Systems*. Springer–Verlag, New York (2004)