STAR Modeling of Pulmonary Tuberculosis Delay-Time in Diagnosis

Bruno de Sousa, Dulce Gomes, Patrícia A. Filipe, Cristiana Areias, Teodoro Briz and Carla Nunes

Abstract Understanding what characterizes patients who suffer great delays in diagnosis of pulmonary tuberculosis is of great importance when establishing screening strategies to better control TB. Greater delays in diagnosis imply a higher chance for susceptible individuals to become infected by a *bacilliferous* patient. A Structured Additive Regression model is attempted in this study in order to potentially contribute to a better characterization of *bacilliferous* prevalence in Portugal. The main findings suggest the existence of significant regional differences in Portugal, with the factor of being female and/or consuming alcohol contributing to increased delay-time in diagnosis, while being an inmate and/or being diagnosed with HIV are factors that increase the chance of an earlier diagnosis of pulmonary TB. A decrease in 2010 to 77% on treatment success in Portugal underlines the importance of conducting more research aimed at better TB control strategies.

Key words: STAR Model, Pulmonary Tuberculosis, Delay-Time, Survival Analysis, Spatial Analysis

Bruno de Sousa
CMDT, UEI-SPIB, *Instituto de Higiene e Medicina Tropical – Universidade Nova de Lisboa*, Portugal, e-mail: bruno.desousa@ihmt.unl.pt

Dulce Gomes and Patrícia A. Filipe
CIMA/UE, *Escola de Ciência e Tecnologia - Universidade de Évora*, Portugal, e-mail: dmog@uevora.pt, pasf@uevora.pt

Cristiana Areias
Escola Nacional de Saúde Pública – Universidade Nova de Lisboa, Portugal, e-mail: c.areias@ensp.unl.pt

Teodoro Briz and Carla Nunes
CMDT, *Escola Nacional de Saúde Pública – Universidade Nova de Lisboa*, Portugal, e-mail: tshb@ensp.unl.pt, CNunes@ensp.unl.pt

1 Introduction

Although many studies have strongly indicated that tuberculosis can be controlled in almost any socio-economical reality [5, 11, 9], it remains a struggle to successfully control TB when faced with the presence of an epidemic HIV infection [6]. The Tuberculosis Programme from the European Center for Disease Prevention and Control (ECDC) recognizes that improvements have been made in Tuberculosis (TB) prevention, but still considers it a threat to human health both world-wide and in Europe. Tuberculosis is currently classified as a re-emerging disease of European importance, with Portugal in 2008 still reporting a notification rates higher than 20 per 100,000 (21 per 100,000 in 2011 [3]), together with Romania, Lithuania, Latvia, Bulgaria, Estonia and Poland [2]. In [3], Portugal, together with other European countries, has a sex ratio of men to women of 2:1, but with a tendency to become more subtle in the future.

The HIV epidemic undermines the control of Tuberculosis. Although the quality and completeness of country data on TB/HIV co-infection vary greatly, out of the eight countries that reported complete data in 2008 with co-infection rates between 0 to 14.6%, Portugal has one of the highest proportions of co-infections cases, together with Estonia and Malta. Nevertheless, Portugal, Iceland and Slovakia achieved the target of a treatment success rate of 85% or higher set by the Stop TB Partnership. Among the 22 studied countries, the successful outcome among previously untreated culture-positive pulmonary TB cases in 2007 was 79.5% [2]. Unfortunately, latest data shows a decrease in 2010 to 77% on treatment success rate in Portugal [3].

Hornick in 2008 [7] reports that data from the Centers for Disease Control and Prevention (CDC) show that approximately 21-23% of individuals coming into close contact with patients suffering from infectious tuberculosis themselves become infected. A multitude of factors could be attributed to the increased risk of an individual contracting TB, as well as of disseminating it if already ill, pulmonary and contagious. With this in mind, the current study explores the effect of some of these factors on the delay-time in diagnosis of Pulmonary TB in Portugal. The factors considered were the addictive consumption of alcohol, drugs, smoking status, as well as the sex, age, number of previous treatments and HIV status of the individual. A structured additive regression (STAR) model was fit to the data in order to explore possible spatial correlations that can arise from the individual's municipality of residence together with the risk factors and other environmental variables, such as being an inmate, homeless or living in a risk area. Living in a risk area was a variable determined from a previously study by Nunes *et al.* [10] where geographical areas were classified as high/low risk areas for contracting Pulmonary TB.

Structured Additive Regression (STAR) models [4] is the class of complex regression models chosen in this study since it allows to take into account a multitude of covariates while exploring possible spatial and temporal correlations. In particular, a Structured Hazard Regression model is applied relaxing the strong condition of proportional hazards in the Cox model [1].

The material and methods are presented in sections 2, followed by the main results (section 3) and a final discussion in section 4.

2 Material and Methods

The database used was provided by two official sources, namely, the National Program for Tuberculosis Control (Pulmonary Tuberculosis notified cases between 2000–2009) and Statistics Portugal – INE (population data). The information provided include lifestyle characteristics of the individuals (alcohol, smoking, drugs), characteristics inherent to the individual (sex, age, number of treatments, new case, HIV) and environmental variables (municipality of residence, being an inmate, homeless, living in a risk area). The delay-time variable will be the focus of our analysis in section 3, representing the time between the first symptoms and the diagnosis of Pulmonary Tuberculosis. Table 1 contains a full description of all variables that will be pursued in the regression analysis performed in Section 4.

Table 1 Description of variables of the Pulmonary Tuberculosis notified cases database

Variable	Description
DelayTime	Time between the first symptoms and the diagnosis of Tuberculosis
Municipality	Municipality where the individual lives (278 municipalities, excluding the autonomous regions of Azores and Madeira)
Sex	Gender of the individual with categories ”male” (= 0) and ”female” (= 1)
Age	Age of the individual in years
Ntreatments	Number of treatments before present diagnosis
Alcohol	Whether an individual consumes alcohol (1 = Yes and 0 = No)
Smoking	Whether an individual smokes (1 = Yes and 0 = No)
Drugs	Whether an individual consumes drugs (1 = Yes and 0 = No)
Inmate	Whether an individual is an inmate (1 = Yes and 0 = No)
Homeless	Whether an individual is homeless (1 = Yes and 0 = No)
HIV	Whether an individual has HIV (1 = Yes and 0 = No)
RiskArea	Whether an individual lives in a risk area (1 = Yes and 0 = No)
NewCase	Whether an individual is a new diagnosed case (1 = Yes and 0 = No)

The results of our analysis will be based on $n = 19,856$ complete notified cases, i.e. the cases for which we have all the information regarding the variables defined in Table 1. This represents 56% of the original database with a delay-time less or equal to 365 days.

A brief descriptive comparison analysis between the data analyzed in this study and the notified cases that were omitted due to missing values was performed. Among the three variables for which we had information for all the notified cases, namely, age, delay-time in diagnosis and sex, the differences between the two groups were very slight, with both groups following the same patterns of behavior.

Survival time and censoring can be modeled through the Cox Proportional Hazards model proposed by David Cox in 1972 [1]. In this model the hazard rate, $\lambda(t|u) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t, u)$, can be interpreted as the instantaneous rate of an event in the interval $[t, t + \Delta t]$, given survival up to time t.

The main goal of survival regression is to describe the influences of covariates u through a regression model for the hazard rate. In the Cox proportional hazards model, the hazard rate is assumed to have a multiplicative structured of the form

$$\lambda(t|u) = \lambda_0(t) exp(u'\gamma) \tag{1}$$

where $\lambda_0(t)$ is an unspecific baseline hazard rate and $u'\gamma$ is a linear predictor formed of (time-constant) covariates u and regression coefficients γ.

Because the ratio between the hazard rates for two individuals with covariates vectors u_1 and u_2 is independent of t, the Cox model (1) is called a proportional hazards model. This assumption is not always present in real life data and needs to be checked for the model to be applied. To account for nonproportional hazards, nonstandard covariate effects, and spatial dimension, the classical Cox model is extended to a nonparametric structured hazard rate model [8] defined as $\lambda_i(t) = exp(\eta_i(t)), i = 1, ..., n,$, with the structured additive predictor defined as:

$$\eta_i(t) = u_i(t)'\gamma + g_0(t) + \sum_{k=1}^{K} g_k(t) w_{ik}(t) + \sum_{j=1}^{J} f_j(\nu_{ij}(t)) \tag{2}$$

where $g_0(t) = \log(\lambda_0(t))$ is the log-baseline hazard, $g_k(t)$ represents time-varying effects of covariates $w_{ik}(t)$, $f_j(\nu_{ij}(t))$ are non linear effects of different types of generic covariates and $u_i(t)'\gamma$ corresponds to effects of parametric covariates.

3 Results

For the hazard rate $\lambda(t) = exp(\eta(t))$ we choose the geoadditive predictor

$$\eta(t) = g_0(t) + f_1(Age) + f_2(Ntreat) + f_{str}(Munic) + f_{unstr}(Munic) + u(t)'\gamma \tag{3}$$

where $g_0(t)$ denotes the log-baseline hazard rate, f_1, f_2 are functions of the covariates age and number of treatments. Functions f_{str} and f_{unstr} models the global and local spatial effects, respectively, based on the municipality where an individual lives. The fixed effects of the numerous categorical covariates (Table 1) are represented by $u(t)$.

We conducted a municipality level analysis, with a Markov random field prior for the spatial effect. The estimates for the log-baseline g_0 and the nonparametric effects f_j are shown in the next two figures. The log-baseline, Figure 1 (a), shows a steep increase until approximately 44 days, followed by an alternate period between positive and negative effects (approximately constant) until around 340 days. At the end of the observation period, there is a strong increase in g_0. However, only 69 individuals had a delay-time in diagnosis more than 340 days and, therefore, this increase should not be over-interpreted.

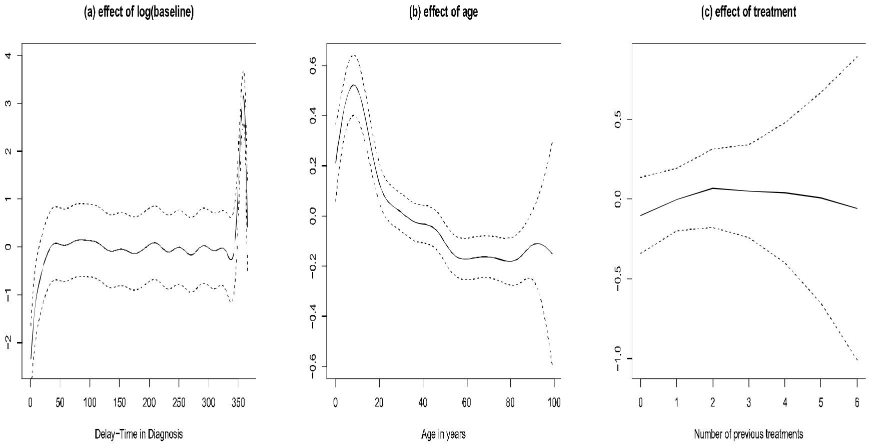

Fig. 1 Municipality-level analysis: Posterior mode estimates of the effects of the log-baseline (a), age (b), and the number of previous treatments (c), together with pointwise 95% credible intervals (dashed lines).

From Figure 1 there is a non-linear effect for both effects f_1 and f_2 in equation (3), regarding age and number of previous treatments before the current diagnosis, with a clear stronger effect in the case of the age effect. There is an increased risk of the event for younger ages ($<$ 8 years), with a steep decrease until age 20, followed by moderate decreases and periods of almost flatness at 35 and 55 years of age. In other words, younger people have an higher chance of being diagnosed earlier, with a clear decrease as people grow older (Figure 1(b)). After 80 years of age we see a slight increase in the chance of an earlier diagnosis, but again this is the result of only 531 individuals and should not be over-interpreted.

The effect of the number of previous treatments before the current diagnosis is almost constant with a slight increase for those who had at least 2 treatments before. However, since the credible intervals include zero, the influence of the number of previous treatments can be neglected (Figure 1(c)). Also of note is the large bandwidth for values higher than 2. This is due to the small number of cases of people with more than 2 previous treatments (58 cases).

Looking at the estimated global spatial effects in the left panel of Figure 2(a), we find an unclear pattern in terms of spatial effects. Municipalities with higher risk of a delayed diagnosis seem to occur in the south litoral, interior Alentejo, the greater Lisbon area, center interior and, with a milder effect, in the north of Portugal. This structure is confirmed by the significance map of Figure 2(b), where black denotes districts with strictly negative credible intervals and white denotes districts with strictly positive credible intervals, i.e. representing a gray scale of municipalities that clearly contribute to a higher risk of a delayed diagnosis (in black) as compared to the ones which contribute to decreasing that risk (in white).

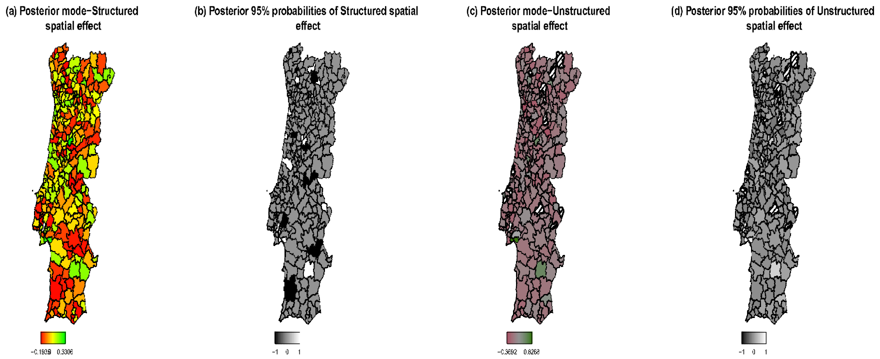
(a) Posterior mode−Structured spatial effect
(b) Posterior 95% probabilities of Structured spatial effect
(c) Posterior mode−Unstructured spatial effect
(d) Posterior 95% probabilities of Unstructured spatial effect
−0.1929 0.3306
−1 0 1
−0.3692 0.6256
−1 0 1

Fig. 2 Municipality-level analysis: Estimated global (a) and local (c) spatial effects, and point-wise 95% significance map ((b) and (d), respectively). Black denotes municipalities with strictly negative credible intervals, whereas white denotes municipalities with strictly positive credible intervals.

In terms of the estimated local spatial effects in Figure 2(c), we observe quite an homogeneous map with most of the regions grey and some areas pink, indicating no local effects in general, with some municipalities showing a tendency for delayed diagnosis. This structure is confirmed by the significance map in Figure 2(d), where black denotes districts with strictly negative credible intervals (higher risk of a delayed diagnosis) and white denotes districts with strictly positive credible intervals. Most of the the municipalities with an increased local risk for a delayed diagnosis are in the center region of Portugal.

Table 2 contains the estimates, standard deviations, p-values and 95% credible intervals of the fixed effects of model in equation 3. Being an individual in a risk area, a new case, a homeless person or one consuming drugs does not seem to be a statistically significant factor in the delay-time in diagnosis. Among the statistically significant factors, being female and/or consuming alcohol seem to reduce the risk of the event (being diagnosed), i.e., increasing the delay-time in diagnosis. On the other hand, being an inmate and/or being diagnosed with HIV increase the chances of an earlier diagnosis of TB. Smoking is a borderline non-significant effect, increasing the chances of an earlier diagnosis as well.

In addition, we conducted a district level analysis where the 278 municipalities were classified into 18 districts. The covariates and fixed effects were very similar to the results presented above. Of note was the very clear pattern in terms of spatial effects. Districts in the center and south of Portugal seem to have a higher risk of a delayed diagnosis. In terms of the estimated local spatial effects, Lisbon and Oporto are the districts that emerged as those that contribute to an increase or decrease in

Table 2 Municipality-level analysis: Estimates, standard deviations, p-values and 95% credible intervals of fixed effects.

Variable	Post. Mode	Std. Dev.	p-value	95% Credible Interval	
const	-3.97602	0.407842	7.581e-10	-4.77556	-3.17649
Sex	-0.11317	0.016025	9.59e-08	-0.14458	-0.08175
Alcohol	-0.06112	0.021552	0.005029	-0.10337	-0.01887
Smoking	0.058227	0.032663	0.074183	-0.00581	0.122259
Drugs	-0.02116	0.029855	0.478341	-0.07969	0.037367
Inmate	0.123101	0.057572	0.032286	0.010237	0.235966
Homeless	0.019203	0.057023	0.736068	-0.09259	0.130991
HIV	0.122223	0.023624	7.45e-06	0.075910	0.168535
RiskArea	-0.02179	0.177806	0.902919	-0.37036	0.326780
NewCase	-0.14126	0.161087	0.380796	-0.45706	0.174531

the delay-time in diagnosis, respectively. Neither global nor local spatial effects lead to strictly negative or positive credible intervals.

4 Discussion

The Delay-Time is the time between the appearance of the first symptoms and the diagnosis of a Pulmonary TB case by the health care system. This time period depends on the actions of both the patient and health care. In Portugal, 95% of the TB cases are diagnosed because symptomatic (with cough) patients search health care services [3]. This means that when a patient is diagnosed, he/she has already infected someone, which can lead to an endless endemic state.

Understanding what characterizes those patients who suffer great delays in diagnosis may contribute, for example, to establishing a better screening strategy, and therefore, a decrease of the endemic level in Portugal. Our study suggests that younger people have an higher chance of being diagnosed earlier, with a clear decrease as people grow older. As reported in [3], age patterns have been changing among TB patients in Portugal as well as in other developed countries. Over time, a decrease of incidence in younger individuals (0–44 years) and an increase in older groups (especially 45–74 years) has been observed [3]. Future analysis will focus on these age groups in order to better understand the possible reasons for this behavior.

Municipalities with higher risk of a delayed diagnosis seem to occur in the south litoral, interior Alentejo, the greater Lisbon area, center interior and, with a milder effect, in the north of Portugal. There are mainly three important factors that can clearly contribute to a delay in the diagnosis: (1) The population's lack of knowledge, (2) Inefficient health care services, and (3) Low incidence of the disease, implying a lesser chance of it being a primary diagnosis. Further research is needed in order to determine the possible reasons affecting these municipalities that might explain these results.

Among the factors considered in this study, being female and/or consuming alcohol indicates a tendency for an increasing delay-time in diagnosis, while being an inmate and/or being diagnosed with HIV increases the chance of an earlier diagnosis of pulmonary TB. Smoking is a borderline non-significant effect, increasing the chances of an earlier diagnosis as well. The knowledge of HIV status has been increasing in Portugal, from 59% of missing cases in 2000 to 41% in 2009. In future research the evolution of the results according to year of notification will be considered in order to be able to identify possible biases due to missing data.

With this initial study, we attempted to understand some of the main risk factors that might be responsible for greater delays in diagnosis, a pivotal piece of the puzzle in order to successfully control TB. Further studies are needed in order to clearly understand the problem within a more global perspective and address some of the research questions stated in this discussion.

Acknowledgements This work was financed by FCT/MCTES through the research project PTDC/SAU-SAP/116950/2010.

References

1. Cox, D.R.: Regression models and life tables (with discussion). Journal of the Royal Statistical Society B **34**, 187–220 (1972)
2. ECDC: Progressing towards TB elimination – A follow-up to the Framework Action Plan to Fight Tuberculosis in the European Union. ECDC Special Report (2010) Available in http://www.ecdc.europa.eu/en/publications/Publications/101111_SPR_Progressing_towards_TB_elimination.pdf.
3. DGS: Relatório StopTb2012-Ponto da Situação Epidemiológica e de Desempenho (2012) Available in http://www.portaldasaude.pt/NR/rdonlyres/8E0DFF04-F030-43B4-80EB-A71AD96F3718/0/relatorio_tuberculose_2012.pdf.
4. Fahrmeir. L., Kneib, T., Lang, S.: Penalized structured additive regression for space-time data: a Bayesian perspective. Statistica Sinica **14**, 731–761 (2004)
5. Frieden, T.R., Fujiwara, P.I., Washko, R.M., Hamburg, M.A.: Tuberculosis in New York City–turning the tide. New Engl J Med **333**, 229–233 (1995)
6. Frieden, T.R.: Can tuberculosis be controlled? International Journal of Epidemiology **31**, 894–899 (2002)
7. Hornick, D.B.: Tuberculosis. In: Robert, B.W., MD, MSc (eds.) Public Health & Medicine, 248-257. New york: MacGraw Hill Medical (2008)
8. Kneib, T.: Mixed Model based Inference in Structured Additive Regression. Fakultät für Mathematik, Informatik und Statistik der Ludwig-Maximilians-Universität München, Germany (2006)
9. Marrero, A., Caminero, J.A., Rodriguez, R., Billo, N.E.: Towards elimination of tuberculosis in a low-income country: the experience of Cuba, 1962–97. Thorax **55**, 39–45 (2000)
10. Nunes, C., Briz, T., Gomes, D., Filipe, P.A.: Pulmonary Tuberculosis and HIV/AIDS: joint space-time clustering under an epidemiological perspective. In: Cafarelli, B. (eds.) Proceedings of the Spatial Data Methods for Environmental and Ecological Processes - 2ndEdition, 1–4. Foggia e Gargano (2011)
11. Suarez, P.G., Watt, C.J., Alarcon, E., Portocarrero, J., Zavala, D., Canales, R., Luelmo, F., Espinal, M.A., Dye, C.: The dynamics of tuberculosis in response to 10 years of intensive control effort in Peru International. J Infect Dis **184**, 473–478 (2001)