Incorporating covariates in a flowgraph model for bladder carcinoma

Gregorio Rubio[1⋆], Belén García-Mora[1], Cristina Santamaría[1], and Francisco Santonja[2]

[1] Instituto de Matemática Multidisciplinar. Universitat Politècnica de València
[2] Universitat de València

Abstract. Superficial bladder cancer is a significant public health problem with remaining challenges. Urologists need tools to accurately predict the real evolution of the disease, that help them to improve treatment modalities and follow-up schemes. Multi-state stochastic processes are a convenient framework for modeling the process, and the statistical flowgraph approach is an efficient tool to perform the task. In this paper this approach is improved by incorporating covariates in the model.

Keywords: flowgraph model, bladder carcinoma, Erlang distribution, covariates

1 Introduction

Superficial bladder cancer (non muscle invasive bladder carcinoma, NMI-BC) is a significant public health problem with remaining challenges. The macroscopic tumor is usually removed from the interior of the bladder by means of a surgical endoscopic technique (TUR, transurethral resection). However it has a notable tendency to recur (30-85 %) and less frequently to progress to muscle invasive stages (10-20 %).

Urologists need tools to accurately predict the real evolution of the disease, that help them to improve treatment modalities and follow-up schemes. It is necessary to go beyond EORTC tables [1], and indeed there is active research [2].

Multi-state stochastic processes are a convenient framework for modeling the process, and the statistical flowgraph approach [3] is an efficient tool to perform the task. We successfully tested this methodology in a previous work [4], and in this paper our aim is to incorporate covariates in the model. The inclusion of covariates in the flowgraph analysis is recent [5]. Our model is based on a database obtained from La Fe University Hospital of Valencia (Spain), that records clinical-pathological information from 960 patients, followed between January 1995 and January 2010.

The paper is organized as follows: in section 2 we review a few basic concepts of survival analysis, phase-type distributions and Erlang distributions, needed

⋆ corresponding author.

to build the model. In section 3 we present the essentials of flowgraph models. Section 4 contains the model, and finally, in section 5 some discussion is given.

2 Survival analysis and phase-type distributions

2.1 Survival analysis

Survival analysis techniques deal with the analysis of data related to elapsed time from a well-defined *time-origin* until the occurrence of some particular event or *end-point*.

Let T be the random variable associated with the survival time (time until the ocurrence of the event). The Survival Function is

$$S(t) = \mathrm{P}(\mathrm{T} \geq t) = 1 - F(t)$$

where $F(t)$ is the distribution function of T. It expresses the probability that an individual survives (that is, does not undergo the occurrence of the event) from the time origin to some time beyond t.

The Hazard Function is given by

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathrm{P}(\, t \leq \mathrm{T} < t + \Delta t \,|\, \mathrm{T} \geq t)}{\Delta t},$$

what expresses the hazard rate or the instantaneous event rate.

In survival analysis data are frequently censored [6], what means that the event of interest has not been observed. The time of follow–up of those patients must be taken into account, because it informs us of the fact that the individual was free of event until the present moment.

2.2 Phase-type and Erlang distributions

To modelize lifetimes, mixtures of distribution functions are useful. In this regard, phase-type distributions [7] are quite interesting, because they provide computations with manageable analytical expressions.

The distribution $F(\cdot)$ on $[0, \infty[$ is a phase–type distribution (PH-distribution) with representation (α, T) if it is the distribution of the time until absorption in a Markov process on the states $\{1, \ldots, m, m+1\}$ with generator

$$\begin{pmatrix} T & T^0 \\ 0 & 0 \end{pmatrix},$$

and initial probability vector (α, α_{m+1}) where α is a row m-vector.

The matrix T of order m is non-singular with negative diagonal entries and non-negative off-diagonal entries and satisfies

$$-Te = T^0 \geq 0,$$

where e denotes a column vector with all components equal to one.

The distribution $F(\cdot)$ is given by

$$F(t) = 1 - \alpha \exp(Tt)e, \qquad t \geq 0 \tag{1}$$

and the density by

$$f(t) = \alpha \exp(Tt)T^0. \tag{2}$$

The survival function is

$$S(t) = \alpha \exp(Tt)e \tag{3}$$

and the hazard function is given by

$$h(t) = \frac{\alpha \exp(Tt)T^0}{\alpha \exp(Tt)e}.$$

The Laplace transform is

$$f(s) = \alpha_{m+1} + \alpha(sI - T)^{-1}T^0, \text{for} Re(s) > 0. \tag{4}$$

A particular case of phase-type distribution is the Erlang distribution, which is the basis of our model. An Erlang distribution $E[r, \lambda]$ has a representation (α, T) as a phase-type distribution [8]:

$$\alpha = (1, 0, \ldots, 0)_{1\times r}$$

$$T = \begin{pmatrix} -\lambda & \lambda & & & \\ & -\lambda & \lambda & & \\ & & \ddots & \ddots & \\ & & & -\lambda & \lambda \\ & & & & -\lambda \end{pmatrix}_{r\times r}$$

Phase-type distributions are a closed class for finite mixtures, and form a class weakly dense in the class of general distributions defined on the positive real line. A finite mixture of Erlangs distributions is therefore a phase-type distribution. We utilized the class of mixtures of three Erlang distributions studied in [9]. The distribution function of the elements in this class is given by the expression

$$G(t) = p_1F_1(t) + p_2F_2(t) + p_3F_3(t), \tag{5}$$

with $p_1 + p_2 + p_3 = 1, p_i > 0, i = 1, 2, 3$.

Let us denote the three Erlangs by $E[r_1, \mu_1], E[r_1, \mu_1], E[r_1, \mu_1]$, with $\mu_i > 0$ and r_i a positive integer, $i = 1, 2, 3$. Based on [9], we will consider the particular case with $r_1 = 1, r_2 = 3, r_3 = 5$. Its representation as phase-type distribution is (α, T) where

$$\alpha = (p_1 \quad p_2 \quad 0 \quad 0 \quad p_3 \quad 0 \quad 0 \quad 0 \quad 0) \tag{6}$$

$$T = \begin{pmatrix} -\mu_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\mu_2 & \mu_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\mu_2 & \mu_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\mu_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\mu_3 & \mu_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\mu_3 & \mu_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\mu_3 & \mu_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\mu_3 & \mu_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\mu_3 \end{pmatrix} \tag{7}$$

3 Flowgraph models

A flowgraph model is a graphical representation of a multistate model, that consists of directed line segments (*branches*) connecting the states. The branches are labeled with *transmittances*, that are the multiplication of the transition probability p_{ij} from state i to state j and an integral transform. This transform can be a characteristic function (CF), a moment generating function (MGF), a Laplace transform (LT), or even an empirical transform [5], [10].

In this paper we deal with the flowgraph of Figure 1, that represents the three-state illness-death model.

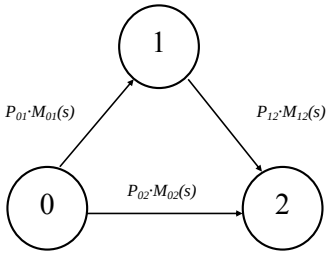

Fig. 1. Three-state illness-death model.

Transmittances are combined according to a systematic procedure (see [3], section 2.5), in order to compute the transforms for the transitions of interest. To manage the graph in Figure 1 only the following rules are needed:
1) The transmittance of transitions in series is the product of the series transmittances.
2) The transmittance of transitions in parallel is the sum of the parallel transmittances.
By applying these rules, the integral transform of the transition of interest is computed. The final step is to invert this transform to recover the probability density function (PDF) of the transition.

4 A flowgraph model for bladder carcinoma with covariates

4.1 Data

The database contains detailed information of patients with NMIBC, obtained from La Fe University Hospital of Valencia (Spain). Data were collected on 960 postoperative patients, between January 1995 and January 2010, with patient characteristics and pathological details. We will distinguish between clinical characteristics of the tumor (number of tumors and size of the tumoral mass), and pathological characteristics (grade and stage). The *stage* of the bladder tumors is classified in Ta and T1 (corresponding to *superficial* bladder tumors according to TNM system classification [11]). *Grade* is categorized from G1 to G3 (from low aggressive to highly aggressive) according to the WHO (World Health Organization [12]). These clinicopathological data were collected when the primary NMIBC was diagnosed and a transurethral resection of the bladder tumor (TURB) was conducted. Also, the variables *Sex* and *Age* were collected at the moment of the *TURB*. *Number* was classified in two levels: one and multiple tumors. *Size* has also two levels: less or equal to 3 cm and more than 3 cm. Characteristics of patients are provided in Table 1.

Patients were followed up every three months during the two first years, every six months between the second and fifth years following diagnosis, and then annually with a minimum follow-up time of ten years. The mean follow–up time for the entire cohort was 3 years and 11 months. The data record several recurrence times. This means that some patients have no recurrence at all, some have one or more recurrences, and some have progression (directly of after some recurrence). In our model we have considered progressions and one recurrence. As stated above, 434 patients underwent a recurrence, 24 a progression, and 499 had censored times. Then, 63 patients were lost. From the remaining 371 patients, 17 underwent a progression. Times of the remaining 354 patients were considered censored.

The joint evolution of the two processes (*recurrence process* and *progression process*) was modelled by means of a non-parametric penalized likelihood method for estimating hazard functions in a general joint frailty model for recurrent events and terminal events in [13]. In this work, four variables (age, grade, number and size of the tumor) were obtained as significant covariates in the recurrence process so these variables are used to fit transition 0-1. In the progression process three variables were obtained as significant: age, stage and grade, these variables are used to fit transitions 0-2 and 1-2.

4.2 Flowgraph model

In [4] parametric distributions were computed for the flowgraph of Figure 1, in the context of our model for bladder carcinoma. State 0 corresponds to the patient free of disease, after the TUR of the primary tumor. State 1 is the first recurrence, and state 2 is progression. Time is given in years.

Variable	patients	%
Sex		
Men	838	87.3
Women	122	12.7
Age		
$\leq$ 66 years	429	44.7
$>$ 66 years	464	48.3
Missing	67	7
Stage		
Ta	287	29.9
T1	650	67.7
Missing	23	2.4
Grade		
G1	373	38.9
G2	392	40.8
G3	171	17.8
Missing	24	2.5
Number		
One	577	60.1
Two or more	168	17.5
Missing	215	22.4
Size		
$\leq$ 3 cm	609	63.4
$>$ 3 cm	222	23.1
Missing	129	13.4

Table 1. Patient and tumor characteristics.

Once the parametric distributions have been computed, the Laplace transforms are easily calculated from (4). Then we compute the Laplace transform relevant to the transitions of interest, applying the above rules. Our interest is to model the overall risk of progression. So we aim to find the probability distribution of time to reach state 2 for the first time starting in state 0, irrespective of the path that was taken. That is to say, the first passage distribution of going from disease free to muscle invasive stages.

The application of rules 1 and 2 is the same as in the case without covariates. So the Laplace transform would be given by:

$$LT(s) = p_{01}p_{12}LT_{01}(s)LT_{12}(s) + p_{02}LT_{02}(s).$$

However, it must be taken into account that our flowgraph is actually part of a more general graph that would model the disease process, see Figure 2. Passage from state 0 to state 2 is not certain to occur: a patient may only suffer

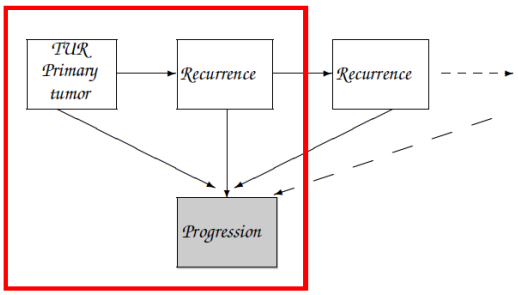

Fig. 2. Recurrence – progression process.

recurrences, or even no recurrence. The probability of taking the considered path is $p_{01}p_{12} + p_{02}$, and we must divide the preceding $LT(s)$ by this probability to obtain the true Laplace transform [3, pag. 19]

$$LT(s) = \frac{p_{01}p_{12}LT_{01}(s)LT_{12}(s) + p_{02}LT_{02}(s)}{p_{01}p_{12} + p_{02}} \tag{8}$$

Probabilities p_{ij} are assigned from estimations based on our data. They simply consist of the ratios between the number of progressions or recurrences and the number of patients who could undergo the relevant transition. Calculations are quite sensitive to these values. We tried with the current and also previous database. The best results were obtained taking $p_{01} = 0.3967742$, $p_{02} = 0.02507837$ and $p_{12} = 0.03252033$.

The final step is to invert these transforms to obtain PDFs, for which we use an inversion algorithm called EULER, developed by Abate and Whitt [14], in the version provided by [15].

4.3 Incorporating covariates

The starting point is the parametric distribution for each transition, obtained without taking into account covariates. They are of the form (6)-(7), with different parameters p_i and μ_i. The approach of [16] suggested us to incorporate covariates by multiplying each μ_i by $\exp(XB^T)$, where X is a vector with the covariates of a patient and B a vector of coefficients. Thus the influence of covariates on the distributions will consist in modulate their shape through the variation of the values of these parameters in the phase type distributions.

From expression (2) the PDFs are easily obtained, and we can make up the likelihood function. The general expression is provided in [5], from which the suitable function for our model is $L = L_0 L_1$, where L_0 and L_1 are computed as follows.

Let S_{jk} be the set of observation indices in transition j-k, and S_{j^*} the index set of observations censored in state j. Let f_{jk} the waiting time density associated with transition j-k, and $t_{jk,i}$ the observed time for patient i. Then

$$L_0 = \left[\prod_{k=1}^{2} \prod_{i \in S_{0k}} p_{0k} f_{0k}(t_{0k,i})\right] \times \left[\prod_{i \in S_{0*}} (1 - F_0(t^*_{0,i})\right]$$

where $t^*_{0,i}$ is the observed censoring time in state 0 for the observation i and F_0 is the cumulative distribution function corresponding to the mixture density

$$f_0(t^*_{0,i}) = \sum_{k=1}^{2} p_{0k} f_{0k}(t^*_{0,i}).$$

And, with the obvious meaning for $t^*_{1,i}$

$$L_1 = \prod_{i \in S_{12}} p_{12} f_{12}(t_{12,i}) \times \prod_{i \in S_{1*}} (1 - F_1(t^*_{1,i})$$

where F_1 is the cumulative distribution function corresponding to

$$f_1(t^*_{1,i}) = p_{12} f_{12}(t^*_{1,i}).$$

The parameter values are obtained by minimizing $-logL$, using as seeds the zero value for each parameter (which corresponds to the no covariate case).

Therefore we can perform the distribution function and all the other functions in each transition for any group of patients with particular values of covariates. For instance, in Figure 3 the empirical and model distributions corresponding to transition 0-1 are shown for the group of patients with the following initial characteristics: one tumor, stage T1, grade G2, smaller than 3 cm, and more than 66 years old. Then the Laplace transform for the transition that we are interested in, is computed from equation (8). From this function we obtain the survival function (with regard to progression) for this particular group. The plot is shown in Figure 4, and jointly with the empirical function in Figure 5.

All computations were made in R [17]. We used Stats [18], expm [19], Matrix [20] and survival [21] packages.

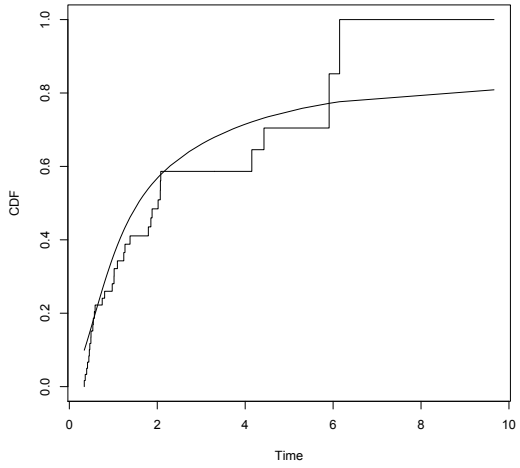

Fig. 3. Erlang mixture (*smooth line*) and empirical distribution (*step function*) for transition 01.

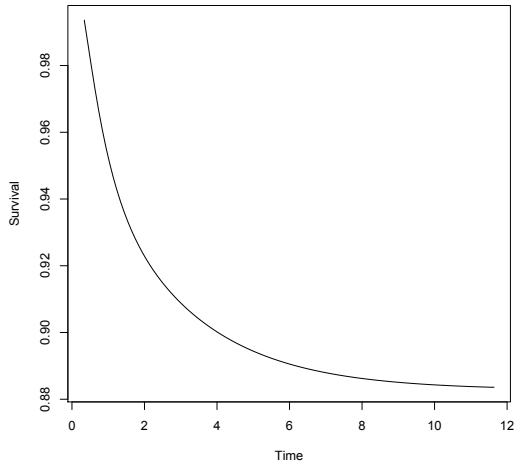

Fig. 4. Survival function obtained from the model.

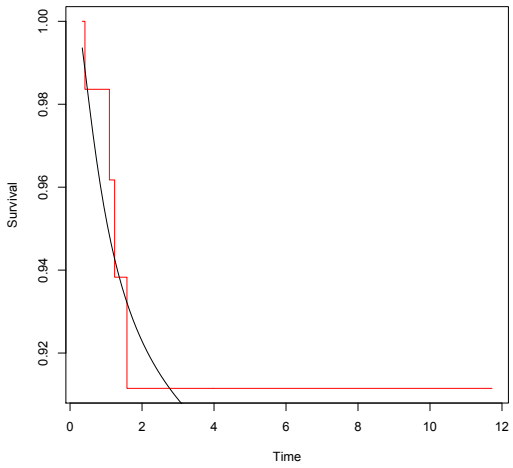

Fig. 5. Survival function model (*smooth line*) and empirical survival function (*step function*).

5 Discussion

The Erlang distributions approach let us incorporate covariates in a relatively simple manner. Computations are quite tractable combined with the flowgraph methodology. The next step in order to a successful modelling would be to take into account multiple recurrences. This question has been discussed in [5], but in our case there is an important difficulty: the succesive recurrences in the same individual are not actually independent events, and therefore the approaches from [5] can not be fully taken into account.

Acknowledgments. This study has been funded by Vicerrectorado de Investigación de la Universitat Politècnica de València. Reference 2406.

The authors thank Dave Collins for his support and specially his EULER code in R.

References

1. Sylvester, R. J., van der Meijden, A. P., Oosterlinck, W., Witjes, J. A., Bouffioux C., Denis, L., Newling, D. W., and Kurth, K. Predicting recurrence and progression in individual patients with stage ta t1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. *European Urology*, 2006; 49:475–7

2. Bedeir Ali-El-Dein, Prasanna Sooriakumaran, Quoc-Dien Trinh, Tamer S. Barakat, Adel Nabeeh and El-Housseiny I. Ibrahiem Construction of predictive models for recurrence and progression in >1000 patients with non-muscle-invasive bladder cancer (NMIBC) from a single centre. *BJU Int*, 2013 Jun;111(8):E331-41
3. Huzurbazar, A. Flowgraph Models for Multistate Time-To-Event Data. New York: Wiley, 2005
4. Rubio, G., García-Mora, B., Santamaría, C. and Pontones, J.L. A flowgraph model for bladder carcinoma. *International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2013)*, March 18-20, Granada, Spain
5. Huzurbazar, A. and Williams, B. Incorporating Covariates in Flowgraph Models: Applications to Recurrent Event Data. *Technometrics*, 2010; 52(2), 198-208
6. Klein, J.P., Moeschberger, M.L.: Survival Analysis. Techniques for Censored and Truncated Data, Second Edition. Springer, New York (2003)
7. Neuts, M. F. Matrix Geometric Solutions in Stocastic Models. An Algoritmic Approach, The Johns Hopkins University Press, Baltimore (1981)
8. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modeling, SIAM, Philadelphia (1999)
9. Pérez-Ocón, R., Segovia. M.C.: Modeling lifetimes using phase-type distributions. In: Aven, T., Vinnem, J.E. (eds.) Risk, Reliability and Societal Safety, Proceedings of the European Safety and Reliability Conference 2007 (ESREL 2007), Taylor & Francis,3rd edition (2007)
10. Collins, D.H., Huzurbazar, A.V.: System reliabyility and safety assessment using non-parametric flowgraph models. Proc. IMechE Part O: J. Risk and Reliability. 222, 667–674 (2008)
11. UICC International Union Against Cancer, TNM Classification of Malignant Tumours, (Edited by P.Hermanek and L.H. Sobin), pp. 135-137, Springer-Verlag, Berlin (1998)
12. World Health Organization,Histological typing of urinary bladder tumours, International Classification of Tumours, Volumen 10, Geneva (1999),
13. S. Luján. Modelización matemática de la multirrecidiva y heterogeneidad individual para el cálculo del riesgo biológico de recidiva y progresión del tumor vesical no músculo invasivo, PhD Thesis, Universitat de València, Valencia (2012)
14. Abate, J., Whitt, W.: The Fourier-series method for inverting transforms of probability distributions. Queueing Syst. 10(1), 5-88 (1992)
15. Collins, D.H., Huzurbazar, A.V.: Prognostic models based on statistical flowgraphs. Appl. Stochastic Models Bus. Ind. 28, 141–151 (2012)
16. Faddy, M.J., Graves, N., and Pettitt, A.N.: Modeling length of stay in hospital and other right skewed data: comparison of phase-type, gamma and log-normal distributions. Value in Health. 12, 309–314 (2009)
17. R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2013) `http://www.R-project.org/`
18. The R Stats Package `http://www.R-project.org`
19. Goulet, V., Dutang, C., Maechler, M., Firth, D., Shapira, M., Stadelmann, M.: expm: Matrix exponential `http://www.R-project.org`
20. Bates, D., Maechler, M.: Matrix: Sparse and Dense Matrix Classes and Methods `http://Matrix.R-forge.R-project.org`
21. Therneau, T.: Survival: Survival analysis, including penalised likelihood `http://r-forge.r-project.org`