

5-1-2003

# Homogeneous Markov Processes For Breast Cancer Analysis

Ricardo Ocaña-Rilola

*Escuela Andaluza de Salud Pública, Granada, Spain, ricardo@easp.es*


Emilio Sanchez-Cantalejo

*Escuela Andaluza de Salud Pública*

Carmen Martinez-Garcia

*Escuela Andaluza de Salud Pública*

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

## Recommended Citation

Ocaña-Rilola, Ricardo; Sanchez-Cantalejo, Emilio; and Martinez-Garcia, Carmen (2003) "Homogeneous Markov Processes For Breast Cancer Analysis," *Journal of Modern Applied Statistical Methods*: Vol. 2 : Iss. 1 , Article 20.

DOI: 10.22237/jmasm/1051748400

Available at: <http://digitalcommons.wayne.edu/jmasm/vol2/iss1/20>

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.

---

# Homogeneous Markov Processes For Breast Cancer Analysis

## **Cover Page Footnote**

The authors would like to thank Dr. Jacques Estève and Angela Maldonado García.

## Homogeneous Markov Processes For Breast Cancer Analysis

Ricardo Ocaña-Riola   Emilio Sanchez-Cantalejo   Carmen Martinez-Garcia

Escuela Andaluza de Salud Pública  
Granada (Spain)

---

Sometimes, the introduction of covariates in stochastic processes is required to study their effect on disease history events. However these types of models increase the complexity of analysis, even for simpler processes, and standard software to analyse stochastic processes is limited. In this paper, a method for fitting homogeneous Markov models with covariates is proposed for analysing breast cancer data. Specific software for this purpose has been implemented.

Key words: Stochastic processes, Markov processes, cancer, covariates

---

### Introduction

Multi-state Markov processes have been introduced recently in health sciences in order to study the evolution of patients through different states or stages before death, even in cases where exact transition times are not known (Kay, 1986). This type of model has been mainly applied in AIDS (De Gruttola & Lagakos, 1989; Frydman, 1992; Mariotto et al., 1992), cancer (Kay, 1986), and psychiatric research (Keiding & Andersen, 1989), employing different methodologies depending on the particular conditions of each study. In practice, it is often useful to use a homogeneous Markov process to model disease history events because generally they are easy to interpret and the assumption that the process is homogeneous simplifies the methods used to fit the model.

In multivariate studies, the use of models that incorporate covariates allows analysis of the effect of these variables on the outcome variable. When multi-state models are used, it is also possible to study the effect of these covariates on different transitions between states throughout the patient's disease history.

Some authors have worked on the introduction of covariates in multi-state processes and particularly in homogeneous Markov processes (Kalbfleisch & Lawless, 1985; Pastorello, 1993); however, they mentioned the increased complexity of analysis in this sort of model where an added problem is the shortage of standard software. In spite of these problems, the introduction of covariates in stochastic processes is required to explain the effect of these factors on disease history events.

In this paper we present a breast cancer study where two transient states and a death state have been defined. In this study, observation is continuous, i.e., information on exact transition times between transient states is available; in this context, the main objectives of this paper are:

- a) To propose a method, computationally tractable, to estimate homogeneous Markov models with covariates in continuous time.
- b) To study the evolution of patients diagnosed with breast cancer in Granada province (South of Spain).

---

Correspondence should be sent to Ricardo Ocaña-Riola, Escuela Andaluza de Salud Pública, C/ Cuesta del Observatorio, 4 Apdo de Correos 2070 18080 Granada (Spain). E-mail: ricardo@easp.es. This research was developed at the *Escuela Andaluza de Salud Pública* and financed by grant number IN92-D24255738 from the *Programa Nacional de Formación de Personal Investigador en España* of the *Ministerio de Educación y Ciencia*. The authors would like to thank Dr. Jacques Estève and Angela Maldonado García.

### Methodology

The study was carried out with 241 women with breast cancer diagnosed in 1985-86 who received radical treatment and had a period free of symptoms. The follow-up ended on 31 of December 1990 (Ocaña-Riola, 2002). Data originated from the Granada Cancer Registry (South of Spain).

The variables T, N and Hormonal Status (HS) on the disease history of individuals have been recorded. The definition of T and N was taken from the Classification of Malignant Tumours (Sobin and Wittekind, 1997), where these variables are two components of the TNM system for describing the anatomical extent of disease. Variable M was not considered because there were no patients with distant metastasis. Additional numbers on TNM components indicates the extent of the malignant tumour as follows:

a) T: The extent of primary tumour; T0: No evidence of primary tumour; T1: Tumour 2 cm or less in greatest dimension; T2: Tumour more than 2 cm but not more than 5 cm in greatest dimension; T3: Tumour more than 5 cm in greatest dimension; T4: Tumour of any size with direct extension to chest wall or skin.

b) N: The absence or presence and extent of regional lymph node metastasis; N0: No regional lymph node metastasis; N1: Metastasis to movable ipsilateral axillary nodes(s); N2: Metastasis to ipsilateral axillary node(s) fixed to one another or to other structures; N3: Metastasis to ipsilateral internal mammary lymph node(s).

It is considered to be a three-state Markov model with two transient states and one absorbing (Chiang, 1968). These states are “With symptoms “ (state 1), “Without symptoms “ (state 2) and “Death “ (state 3) where the possible transitions are represented in Figure 1 in appendix.

We consider the transition intensity matrix:

$$Q(x) = \begin{pmatrix} -(q_{12}(x) + q_{13}(x)) & q_{12}(x) & q_{13}(x) \\ q_{21}(x) & -(q_{21}(x) + q_{23}(x)) & q_{23}(x) \\ 0 & 0 & 0 \end{pmatrix}$$

where each transition intensity is dependent on a vector of covariates; that is:

$$q_{ij}(x) = \exp(x \mathbf{b}_{ij}) \quad i \neq j$$

$$q_{ii}(x) = \sum_{j \neq i} \exp(x \mathbf{b}_{ij}) \quad ,$$

where  $x = (x_0, \dots, x_b)$ ,  $x_0 = 1$ , is a vector of covariates and  $\mathbf{b}_{ij} = (\mathbf{b}_{ij0}, \dots, \mathbf{b}_{ijb})$  is a vector of unknown parameters.

In order to estimate the model an approximate method was used (Ocaña-Riola, 2002). The Likelihood Ratio Statistic (LRS) was used in a backward analysis to test the signification of regression parameters (De Groot, 1986). Moreover, the LRS test was used for the goodness of fit of the final model (Kalbfleisch & Lawless, 1985). When the transition intensity matrix is estimated, the estimated transition probability matrix is  $P(u; x) = \exp(Q(x)u)$ ,  $u > 0$ .

### Results

In order to estimate the model, we used a partition of the time using 35 intervals which extent was between 0.002 and 0.260 years (Figure 2 in appendix). Because of shortage of subjects in the groups N2 and N3 (Table 1), the variable N has been transformed in a binary variable as  $N_i = 0$  if  $N=0$  and  $N_i = 1$  if  $N=1, N=2$  or  $N=3$ .

There were not transitions from state 2 to state 3 in Non-menopause patients, however there are some in the Menopause group; if we interpret 2-3 as the transition to other causes of death, the transitions observed in Menopause group could be due to an age effect because older women heavily weight this group. For this reason we propose the following model:

$$q_{23} = \exp(\beta_{230} + \beta_{231} T_2 + \beta_{232} T_3 + \beta_{233} T_4 + \beta_{234} N_1) \text{ if } HS = 1$$

$$q_{23} = 0 \text{ if } HS = 0$$

$$q_{23} = \exp(\beta_{230} + \beta_{231} T_2 + \beta_{232} T_3 + \beta_{233} T_4 + \beta_{234} N_1) \text{ if } HS = 1$$

$$q_{23} = 0 \text{ if } HS = 0$$

where  $T_2, T_3, T_4$  are dummy variables from  $T$  ( $T_1$  is the category of reference).

A backward analysis using LRS test showed that variable N is not statistically significant when T and HS are into the model ( $P=0.482$ ). Besides, there is no evidence ( $P=0.370$ ) against the codification of T in only two categories: patients with a better prognosis (T1 or T2) and patients with a bad prognosis (T3 or T4). Therefore it was considered a new covariable, TR, with value 0 for T1 or T2 and value 1 for T3 or T4. The final model is shown in Table 2. MLE's for transition intensities in different groups of covariates are in Table 3.

Figures 3 and 4 show these transition probabilities by groups of covariates. These graphs show a notable difference between T1-T2 and T3-T4. A LRS test for the goodness of fit of the final model shows that there is no evidence against a homogeneous Markov process ( $p=0.177$ ).

### Conclusion

Multi-state Markov models offer some advantages over traditional survival models for studying disease history events, making it possible to estimate the probability that a subject could be in different states at any time in the future. Homogeneous processes are the simplest of Markov models but in some studies it is possible to find evidence against this sort of model. The absence of homogeneity in time could be the result of the absence of homogeneity between people. In this case, the use of covariates could improve the fit of the model and homogeneous Markov models with covariates are an interesting option.

However, the incorporation of covariates in a stochastic process increases the complexity of analysis, even on simple processes. Because of that and the shortage of standard software to analyse

Markov process with incomplete observations, many researchers refuse to use these multi-state models. In spite of these problems, some authors worked on the inclusion of covariates in a homogeneous Markov process (Andersen, 1988; Pastorello, 1993; Tuma & Robins, 1990). The more used methods are based on the extended Kalbfleisch and Lawless algorithm to incorporate covariates (Kalbfleisch & Lawless, 1985).

In this article we have used a particular partition of the time when observation is continuous. In this situation an approximate method has been proposed in order to introduce covariates and to estimate the intensity matrix in a homogeneous Markov process (Ocaña-Riola, 2002). MLE's obtained from this method are not computationally costly and, in practice, the algorithm converges to very similar estimates of parameters given by other methods when the length of the intervals  $u_k$  tends to be small (Ocaña-Riola, 2002). Moreover, covariates can easily be introduced in the model.

The method proposed here consider only categorical covariates because this is the sort of variables analysed in the breast cancer study. Continuous covariables, as age, could be introduced in the analysis using different categories for them. This idea has been used in some research about stochastic processes and in practice it is the most used (Tuma & Robins, 1980; Pastorello, 1993).

In this breast cancer study, incorporation of variables T, N and Hormonal Status in the model have allowed us to evaluate its effects on disease history. However, covariate information was missing for 36 women not included in the analysis. In general, it is not a good statistical practice to leave out patients with missing values; therefore different statistical methods have been published recently in order to incorporate these patients into the analysis. Some authors have shown that using a Bayesian approach implemented via Markov Chain Monte Carlo it is possible to obtain a suitable regression model for the missing values (Raghuathan & Siscovick, 1996).

Due to the complexity of the Bayesian analysis in a Markov process with covariates, we have not implemented this method. However, it would an interesting research into stochastic processes.

Along these lines, Volinsky et al. (1997) applied Bayesian Model Averaging to the selection of variables in Cox proportional hazard models.

Their investigations into the risk factors for strokes using this model improve the results obtained by traditional stepwise, forward and backward selection methods, which have poor properties (Miller, 1990). Again, the implementation of Bayesian Statistics into Markov processes could yield interesting results, although some theoretical research is needed before using these methods in practice.

In a traditional backward analysis a relationship was found between T and Hormonal Status and the evolution of patients diagnosed with breast cancer. Non-menopausal women with a tumour T1 or T2 have the best prognosis since recurrence probability and death probability are the smallest. In the same way and using traditional survival models, other population base studies have found that both, T and Hormonal Status, are important factors in order to predict survival and recurrence probability in breast cancer (Coebergh et al., 1995). Other analytic factors and hormonal data, not included in this study, could explain to a great extent part of breast cancer survival and recurrence. Vascular and lymphatic invasion of cancer cells, type of histology, age, site of first recurrence, female sex steroid receptors and ploidy measurements have been reported in some articles as prognostic factors for breast cancer recurrence (Blanco G et al., 1990; Murayama et al., 1986). In this way, a prospective study could be interesting in order to analyse the effect of all these variables on the evolution of patients through different states of their disease, obtaining a complete and detailed study on breast cancer history.

In this study, the interpretation of the transition from “without symptoms “ to “death “ is difficult in menopausal women. Older women heavily weight this group and perhaps the effect of age can explain this situation. It might be interesting to consider a fourth state “death from other causes “ in order to know the proportion of patients dying from the direct or indirect consequences of breast cancer but unfortunately this information is rarely available in the Granada Cancer Registry.

From this paper’s findings, it will be possible to estimate the proportions of patients who shall be in each disease state in the future; therefore we will be able to obtain highly relevant information for health planning services. Furthermore, the proposed method can easily be used for other situations in cancer and other disciplines such as

public health, economics, sociological research or medical sciences.

#### References

- Andersen, P. K. (1988). Multistate models in survival analysis: A study of nephropathy and mortality in diabetes. *Stat. Med.*, 7, 661-670.
- Blanco, G., Holli, K., Heikkinen, M., Kallioniemi, O. P., & Taskinen P. (1990). Prognostic factors in recurrent breast cancer. *Br. J. Cancer*, 62, 142-146.
- Chiang, C. L. (1968). *An introduction to stochastic processes and their applications*. New York: John Wiley & Sons.
- Coebergh, J. W., Van der Heijden, L. H., & Janssen, M. L. (1995). *Cancer incidence and survival in the southeast of the Netherlands, 1955-1994*. Eindhoven: IKZ.
- De Groot, M. H. (1986). *Probability and Statistics*. New York: Addison-Wesley.
- De Gruttola, V., & Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics*, 45, 1-11.
- Frydman, H. (1992). A nonparametric estimation procedure for a periodically observed three-state Markov process, with application to AIDS. *J. Roy. Stat. Soc. B Met.*, 54, 853-866.
- Kalbfleisch, J. D., & Lawless, J. F. (1985) The analysis of panel data under a Markov assumption. *J. Am. Stat. Assoc.*, 80, 863-871.
- Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, 42, 855-865.
- Keiding, N., & Andersen, P. K. (1989) Nonparametric estimation of transition intensities and transition probabilities: a case study of a two-state Markov process. *Appl. Statist.*, 38, 319-329.
- Mariotto, A. B., Mariotti, S., Pezzotti, P., Rezza, G., & Verdecchia, A. (1992). Estimation of the Acquired Immunodeficiency Syndrome incubation period in intravenous drug users: A comparison with male homosexual. *Am. J. Epidemiol.*, 135, 428-437.
- Murayama, Y., Mishima, Y., & Ogimura, H. (1986). Determination of discriminatory power of prognostic factors for recurrence of breast cancer. *Cancer Detect. Prev.*, 9, 449-453.
- Miller, A. J. (1990). *Subset selection in regression*. London: Chapman and Hall.
- Ocaña-Riola, R. (2002) Two methods to estimate homogeneous Markov processes. *Journal of Modern Applied Statistical Methods*, 1, 131-138.

Pastorello, S. (1993). La mobilità nel mercato del lavoro: un'analisi econometrica con osservazioni in tempo discreto. *Statistica*, 53, 185-206.

Raghunathan, T. E., & Siscovick, D. S. (1996). A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Appl. Statist.*, 45, 335-352.

Tuma, N. B., & Robins, P. K. (1980). A dynamic model of employment behavior: an application to the Seattle and Denver income maintenance experiments. *Econometrica*, 48, 1031-1052.

Sobin, L. H., & Wittekind, C. H. (1997). *TNM Classification of Malignant Tumours*. New York: John Wiley & Sons.

Volinsky, C. T., Madigan, D., Raftery, A. E., & Kronmal, R. A. (1997) Bayesian Model Averaging in proportional hazard models: assessing the risk of a stroke. *Appl. Statist.*, 46, 433-448.

## Appendix

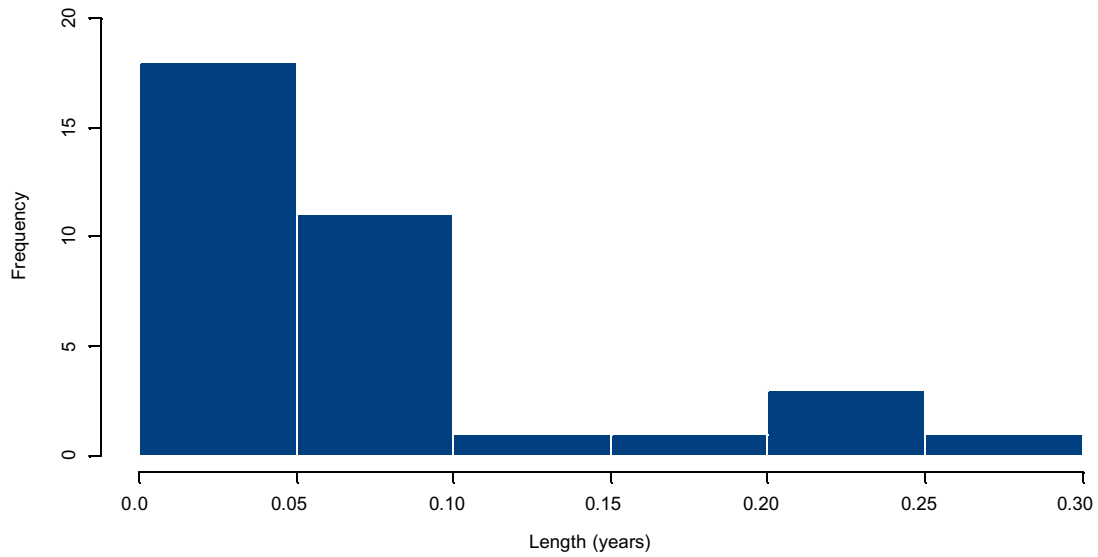
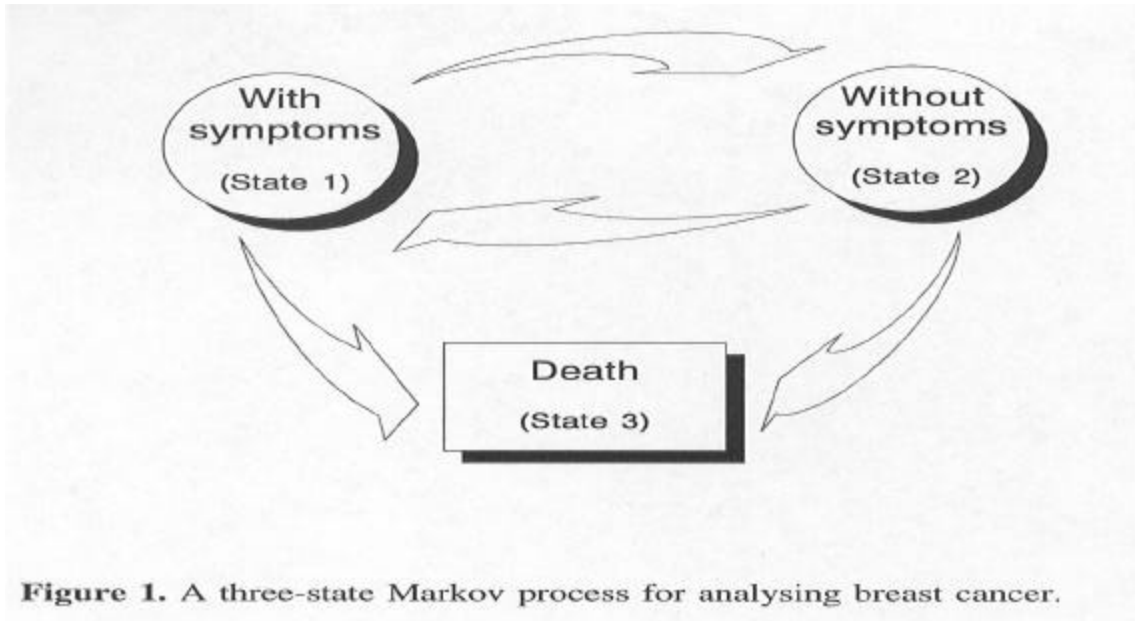


Figure 2. Length of the intervals that give a partition of the follow-up time



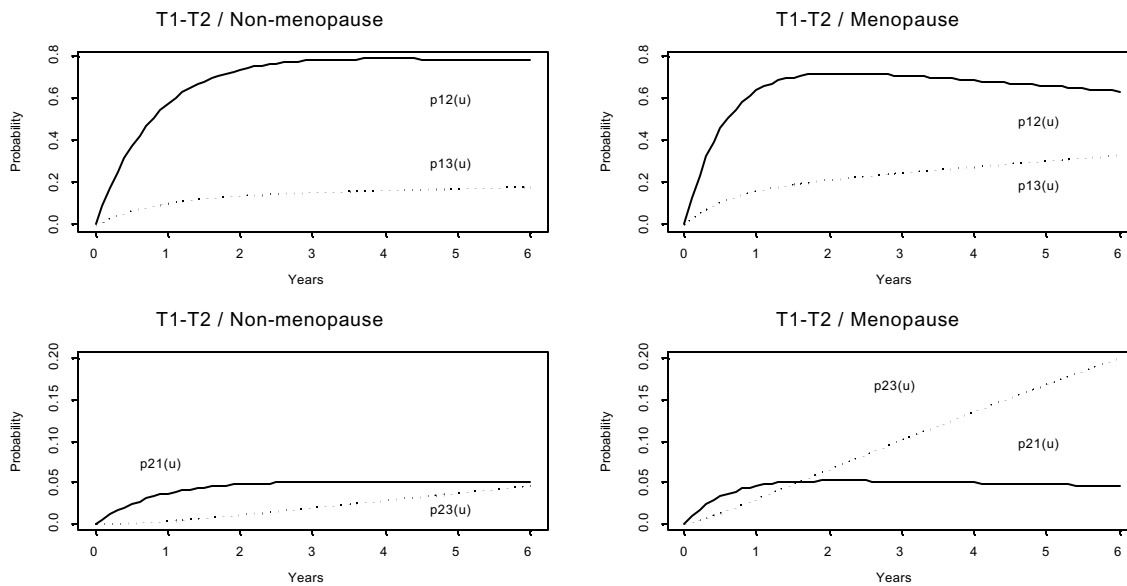


Figure 3. Estimated transition probabilities for T1-T2 and hormonal status

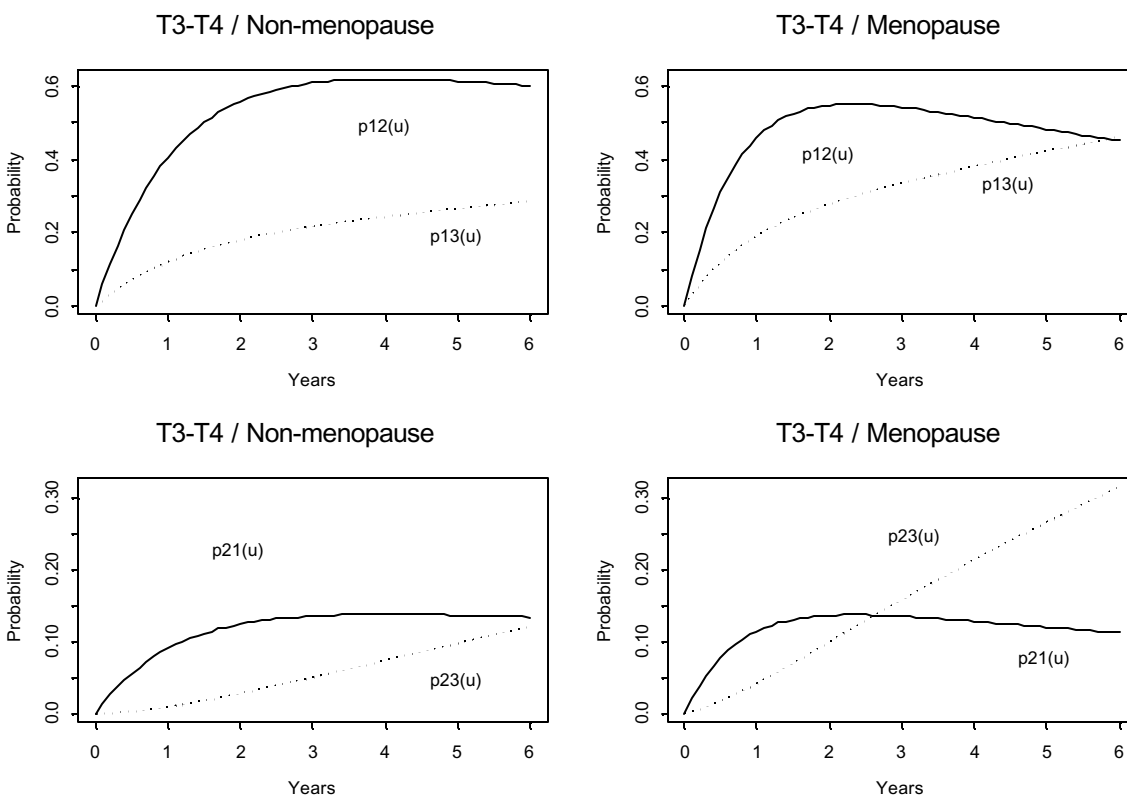


Figure 4. Estimated transition probabilities for T3-T4 and hormonal status

Table 1. Breast cancer data. Granada Cancer Registry, 1985-1986.

	Non-menopause					Menopause				
	N0	N1	N2	N3	Total	N0	N1	N2	N3	Total
T1	15	8	0	0	23	29	6	0	0	35
T2	20	7	1	0	28	41	21	3	0	65
T3	1	5	1	0	7	7	4	1	2	14
T4	2	7	0	0	9	4	15	2	3	24
Total	38	27	2	0	67	81	46	6	5	138

Note: There were 36 patients with missing values.

Table 2. MLE's estimates for breast cancer data (standard error in brackets)

Transition ( $ij$ )	Constant ( $\mathbf{b}_{ij0}$ )	TR ( $\mathbf{b}_{ij1}$ )	Hormonal Status ( $\mathbf{b}_{ij2}$ )
1 - 2	*	-0.4665 (0.0108)	0.3321 (0.0067)
1 - 3	-1.7584 (0.0203)	*	0.5802 (0.0235)
2 - 1	-2.7298 (0.0169)	0.7570 (0.0166)	0.4442 (0.0183)
2 - 3	-3.7965 (0.0219)	*	No included

(\*) Null statistical significance for  $\mathbf{a} = 0.05$

Table 3. Estimated transition intensities for breast cancer data.

T	Hormonal Status	$\hat{q}_{12}$	$\hat{q}_{13}$	$\hat{q}_{21}$	$\hat{q}_{23}$
T1 or T2	Non-menopause	1.0000	0.1723	0.0652	0
T1 or T2	Menopause	1.3939	0.3078	0.1017	0.0224
T3 or T4	Non-menopause	0.6272	0.1723	0.1391	0
T3 or T4	Menopause	0.8742	0.3078	0.2168	0.0224