MODELLISTICA E CALCOLO SCIENTIFICO

MODELING AND SCIENTIFIC COMPUTING

MOX–Report No. 24/2014

Multi-State modelling of repeated hospitalisation and death in patients with Heart Failure: the use of large administrative databases in clinical epidemiology

Ieva, F., Jackson, C.H., Sharples, L.D.

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox@mate.polimi.it   http://mox.polimi.it

Multi-State modelling of repeated hospitalisation and death in patients with Heart Failure: the use of large administrative databases in clinical epidemiology

Francesca Ieva[1], C H. Jackson[2], Linda D. Sharples[3]

June 24, 2014

[1] Department of Mathematics "Federigo Enriques",
Universit degli Studi di Milano, Milano, Italy.
francesca.ieva@unimi.it
[2] Medical Research Council Biostatistics Unit, Cambridge (UK).
chris.jackson@mrc-bsu.cam.ac.uk
[3] Clinical Trials Research Unit, Leeds Institute of Clinical Trials Research, University of Leeds, Leeds (UK).
L.Sharples@leeds.ac.uk

Abstract

In chronic diseases like Heart Failure (HF), the disease course and associated clinical event histories for the patient population vary widely. To improve understanding of the prognosis of patients and enable health-care providers to assess and manage resources, we wish to jointly model disease progression, mortality and their relation with patient characteristics. We show how episodes of hospitalisation for disease-related events, obtained from administrative data, can be used as a surrogate for disease status. We propose flexible multi-state models for serial hospital admissions and death in HF patients, that are able to accommodate important features of disease progression, such as multiple ordered events and competing risks. Markov and semi-Markov models are implemented using freely available software in R. The models were applied to a dataset from the administrative data bank of the Lombardia region in Northern Italy, which included 15,298 patients who had a first hospitalisation ending in 2006 and 4 years of follow up thereafter. This provided estimates of the associations of of age and gender with rates of hospital admission and length of stay in hospital, and estimates of the expected total time spent in hospital. For example, older patients and men were readmitted more frequently, though the total time in hospital was roughly constant with age. We also discuss the relative merits of parametric and semi-parametric multi-state models, and assessment of the Markov assumption.

Keywords: Multi State Models, Heart Failure, Administrative Data, Hospital Admissions, Competing Risks.

1 Introduction

Aging of the population and improved survival of cardiac patients due to modern therapeutic innovations has led to an increasing prevalence of heart failure (HF). Despite improvements in therapy, the mortality rate in patients with HF remains high.[1] The magnitude of the problem of HF is difficult to assess with precision since there is no gold standard for the diagnosis of heart failure, and there has been wide variation in the diagnostic criteria used in different studies.[2] At least six HF scoring systems based upon symptoms and signs have been developed to assess the presence or severity of heart failure. Clinical diagnostic criteria for heart failure have generally included history, physical examination, and chest radiographs (see Mosterd *et al.*,[3] Roger[4] and references therein). Regardless of the definition used, the prevalence of HF and left ventricular dysfunction increases steeply with age (see, for example, Bleumink *et al.*[5]). In general HF is a chronic disease (Chronic Heart Failure — CHF), caused by many conditions that damage the heart muscle, including coronary artery disease, heart attack, cardiomyopathy and conditions that overwork the heart (high blood pressure, valve disease, thyroid disease, kidney disease, diabetes, or heart defects present at birth). In addition, HF can occur in the presence of a combination of these diseases. It is the leading cause of hospitalisation in people older than 65 years. A 2010 update from the American Heart Association (AHA) estimated that there were 5.8 million people with HF in the United States in 2006 (see McMurray *et al.*[6] and Lloyd-Jones *et al.*,[7] among others). There are an estimated 23 million people with HF worldwide. In the Lombardia district of Italy, which provides our motivating example, the HF incidence over the last decade ranged between $25,000$ and $30,000$ cases per year in a population of 9.7 million inhabitants.[8]

In chronic diseases like CHF, clinical interest lies in both the final outcome (death or survival time) and the dynamics of the process itself. To improve understanding of prognosis and for healthcare providers to assess the impact and costs of the disease, a comprehensive model should include both death and non-fatal clinical events. There are several methodological approaches to the modelling of times to multiple events per subject. Castaneda and Bart[9] provide an appraisal of several methods, highlighting that the standard Cox model is not appropriate since observations are not independent. In order to overcome this, they propose the use of marginal and multi-state models using a counting process approach for the joint analysis of survival and time to disease-related hospitalisations, allowing for population average estimates of treatment effects. Several marginal models are adapted in order to account for intra-subject correlation and competing risks. The models differ in the way they define the ”at-risk” population at each time. However in these marginal models it is assumed that all events are identical and can be revisited at any time, with no recognition of the serial nature of consecutive HF-hospitalisations. In their multi-state models, the serial nature of the events is allowed, but hospitalisation and death are treated as the same type of event, which, given their nature and severity, is unacceptable clinically. Thus, a multi-state model that represents multiple ordered events per subject, accounts for competing risks, and distinguishes between death and hospitalisation, is required.

A multi-state model is a stochastic process in which subjects occupy one of a set of discrete states at any time. Multi-state models are convenient for describing longitudinal data and/or repeated events. In Andersen and Keiding[10] a counting process representation is stressed. In medical applications, the states may represent healthy, different severities of disease, or periods in hospital, and transition rates between states may be modelled in terms of covariates. See, for example, Hougaard[11] for a review, and Commenges,[12] Cook,[13] Putter *et al.*,[14] Sommen *et al.*,[15] Sharples and Titman,[16] Duffy *et al.*,[17] Kay,[18] Chen *et al.*,[19] Commenges and Joly[20] for applications to many different diseases. In Sutradhar *et al.*,[21] multi-state models are developed, in order to compare trends in hospitalisations among cancer survivors. Despite the importance of CHF both in terms of incidence and related human and monetary costs (WHO[22] defines the rising incidence and prevalence of chronic diseases as one of the major global concern), there are few examples in the literature of the application of multi-state models to hospitalisation and death from CHF. Postmus *et al.*[23] used a three-state model representing in hospital, out of hospital or death for 1023 patients from a randomized controlled trial with heart failure.

In this study, the impact of CHF is assessed using data from administrative databases, which provide information on the number and times of hospital admissions and time to death (or administrative censoring). Administrative databases play a central role in the evaluation of health-care systems, due to their widespread diffusion and low cost of information. There is increasing agreement among clinical epidemiologists on the validity of disease and intervention registries based on administrative databases (see, for example, Barbieri *et al.*,[24] Wirhenetal[25] and references therein). A key issue is the selection criteria of the observation units: different criteria may result in different estimates of prevalence or incidence of diseases (Saczynski *et al.*[26]). The use of prospective patient management databases is of current interest (see, for example, Macchia *et al.*,[27] Au *et al.*,[28] Aylin *et al.*,[29] Philbin and DiSalvo[30]). The benefits of using these data for health system planning and evaluation are many: they are population based, often combine information from multiple centres, capture real health system use, are longitudinal and are relatively inexpensive to construct and use. In addition, individual health administrative records can be linked to other data (clinical registry, public health, socioeconomic etc.). The validity of this approach is critically dependent on the reliability of the data and the accuracy of disease coding in the administrative records, as shown, for example, by Lee *et al.*[31] and Saczynski *etal.*[26] If search and data linkage strategies are not carried out rigorously, administrative data on hospital admissions can be less complete and exhaustive than data from epidemiological cohort studies and clinical trials. Despite issues surrounding data reliability, and the on-going debate regarding their use in clinical research (see, for example, Quach *et al.*[32]), significant improvements have been achieved in this area in the last decade, and the use of administrative databases in clinical biostatistics has become an accepted practice (see, among others Schultz *et al.*,[33] Muggah *et al.*,[34] Iron *et al.*[35] and references therein).

We propose a multi-state modelling strategy for the joint analysis of outcomes and hospital admissions in CHF patients, whose data come from the administrative database of an Italian regional district (Lombardia). Our aim is to demonstrate a flexible approach that is able to capture important features of disease progression, such as multiple ordered events and the competing risks of death and hospitalisation, in a novel application. We go further than Postmus *et al.*[23] by using multiple states representing subsequent periods spent in and out of hospital, in order to model how the risk of death and further hospitalisation changes through time and with disease progression. Analyses are carried out using freely-available statistical software `R`.[36] Specifically, the `survival`,[37] `mstate`[14] and `msm`[38] packages are used to fit the multi-state models to the data. This work will provide healthcare providers with an effective modelling tool, using hospital admissions to gain insights into the burden of heart failure, how it relates to patient characteristics and how it changes over time.

We describe the data extraction and inclusion criteria in Section 2, and explain the multi-state modelling methods in Section 3. Key results from applying these methods to the Lombardia HF admissions data are presented in Section 4. In Section 5 we end with a discussion of the strengths and challenges of modelling disease progression through administrative data.

2 Study Cohort and Extraction Criteria

Within the Italian health-care regulation system, every hospital admission produces a record in the administrative database. These records are then collected in an data warehouse called SDO (*Scheda di Dimissione Ospedaliera*, i.e., hospital discharge paper) database. The SDO database has been interrogated to identify heart failure episodes and subsequent hospitalisations. In addition, information both on patient (sex, date and place of birth, residence, . . .) and on hospitalisations (date of admission and discharge, diagnoses and procedures, type of admission, type of discharge, vital status at discharge, . . .) over time can be retrieved.

For the current study we used data extracted for the project *"Utilization of Regional Health Service databases for evaluating epidemiology, short- and medium-term outcome, and process indexes in patients hospitalised for heart failure"*. These data include cases of CHF in the administrative data warehouse of Regione Lombardia, the region in the northern part of Italy with capital Milan. The project aims to describe the epidemiology and natural history of HF patients at regional levels, to profile health service utilisation (e.g. hospitalisations, cardiac rehabilitation, diagnostic tests, outpatient visits, etc.) over time, and investigate variation in patient care according to geographic area, socio-demographic characteristics and other clinical variables.

In order to include the vast majority of HF cases, any admission that ended between 2000 and 2010 in Major Diagnostic Category (MDC) 01 (Nervous System), 04 (Respiratory System), and 05 (Circulatory System) in patients resident in the Nothern Italy regional district of Lombardia has been considered. For people who died by the end of the study, the date of death has been obtained through database linkage with the Italian National Registry of deaths. A list of ICD-9-CM codes relating to HF was created as the union of codes from "Heart failure mortality rate" by AHRQ-IQI[39] and from CMS-HCC[40, 41] Model Category 80. From this dataset admissions for HF were identified if any of the six HF diagnosis fields of the SDO were recorded.

Starting from this population, patients whose first admission (*incident event*) ended during 2006 were selected. The number of hospital admissions for HF and the corresponding dates of admission and discharge were recorded over a 4-year follow up (up to December 31th, 2010). Data were anonymised, labelling each patient with an encrypted ID code.

The eligible cohort consisted of $15,856$ patients (corresponding to $36,949$ records). Among these, patients who were younger than 18 years at the first hospitalisation time were excluded (62 pts., corresponding to 182 rows). Among the remaining cohort, we also removed patients admitted and discharged on the same day, i.e., patients whose length of stay (LOS) in hospital was zero (477 pts., corresponding to 2476 rows), or those having long-stay recovery (LOS greater than 180 days, 19 pts., corresponding to 67 rows). Some other pre-processing and cleaning operations were carried out, for example to check coherence in patients' time-line progressions and test for agreement in event indicators. There were no missing data. The final dataset contained records from $15,298$ patients (corresponding to $35,224$ records),

3 Multi-state Models for HF data

3.1 Definitions

To characterise the association between hospital admissions, mortality and patient characteristics, we adopt a multi-state model describing how an individual moves between a series of discrete states in continuous time. Suppose an individual is in state $S(t)$ at time t. The next state to which the individual moves, and the time of the change, are governed by a set of *transition intensities* $q_{rs}(t)$, $r, s = 1, \ldots, R$. The intensity, or *hazard*, represents the instantaneous risk of moving from state r to state s. This may depend on the time t since the start of the process, patient characteristics $\mathbf{z}(t)$, and possibly also the "history" of the process up to that time, $\mathcal{H}_t$: the previous states visited by the individual and the times spent in them. Therefore, for this patient,

$$q_{rs}(t) = \lim_{\delta t \to 0} \mathbb{P}(S(t+\delta t) = s | S(t) = r)/\delta t$$

are then elements of a $R \times R$ matrix $Q(t)$ whose rows sum to zero, so that the diagonal entries are defined by $q_{rr}(t) = -\sum_{r \neq s} q_{rs}(t)$, and $q_{rs}(t) = 0$ if a transition from state r to state s is not allowed.

3.2 Model structure for HF hospitalisation

The 11 states and the 19 permitted transitions in our application are illustrated in Figure 1. Each patient starts in state 1_I, representing the first hospital admission. From there they can either be discharged from hospital, or die in hospital. Once a patient is out of hospital, they can either be admitted again or die, and once in hospital they can either be discharged or die. Death from any cause is included. A maximum of 6 hospital admissions are modelled, and subsequent admissions (but not deaths) are ignored, due to the sparsity of data from individuals with more than 6 admissions (Table 1). Thus "greater than 5 admissions" is considered as a clinically-important "severe" disease state.

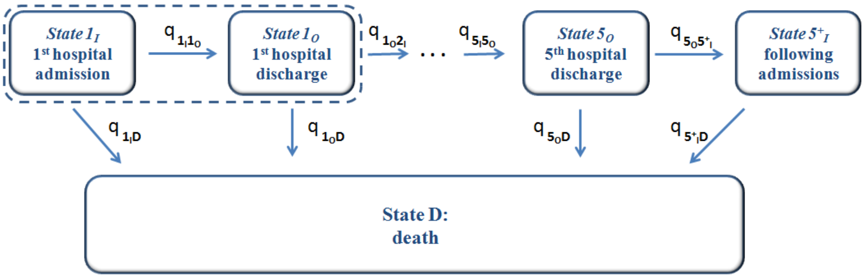


Figure 1: The multi state model, describing jointly the length of stay in hospital, risk of subsequent hospitalisation, mortality rates in and out of hospital, and how these change with increasing numbers of hospital admissions.

Simplifications of this structure are possible. For example, we could have only two living states, representing "in hospital" and "out of hospital", with transitions allowed between them in either direction. This would allow estimation of the in-hospital and out-of-hospital mortality rate, and the average length of stay in hospital, but it would then be awkward to model how these quantities, or the probability of readmission, vary with the previous number of hospital admissions. Alternatively, if length of hospital stay is not of interest, we could omit the "discharge" states, and simply model the times between successive hospital admission dates, jointly with mortality. This would assume, however, that the risk of death does not change when a patient is in hospital. Both of these simplifications were investigated in exploratory work before deciding to use the most flexible structure of Figure 1.

3.3 Data structure and time-to-event modelling

In our application, the state $S(t)$ is known at all times for each patient, since all dates of admission to, or discharge from, hospital, and all dates of death, are known. We label these times $t_1, \dots, t_n$. t_1 is the date of the first hospital admission. If the patient died, t_n is the date of death, otherwise it is the end of follow-up. Any intermediate times $t_2, \dots, t_{n-1}$ represent discharges and subsequent admissions, if they occur.

For each permitted $r \to s$ transition in the multi-state model (19 in our case) there is a corresponding *time-to-event model*, with cause-specific hazard rates defined by $q_{rs}(t)$. To enable estimation of these hazards, the data are expressed as a series of times to events which are potentially censored: $dt_j = t_{j+1} - t_j : j = 1, \dots, n-1$. For a patient who moves into state s at time t_j, their next event at t_{j+1} is defined by the model structure (Figure 1) to be one of a set of competing events $s^*_1, \dots, s^*_{n_s}$.

For example, in state $s = 1_I$ (first hospital admission), the next state must either be $s^*_1 = 1_O$ (first discharge), or death ($s^*_2 = D$) so $n_s = 2$. The time of the event which actually occurs at t_{j+1} is *observed*, and the times of the *competing* events from this set (which have not occurred by this time) are *censored*. Each dt_j contributes an *observed* time to one of the 19 transition-specific models, and a *censored* time to each of the models for the competing events. Therefore, standard tools for survival analysis can be used to estimate the $q_{rs}(t)$, independently for each $r \to s$ transition, from this form of data. Additional software is required to deal with the multi-state structure when processing the data, making predictions (Section 3.4) and presenting results.

We apply two alternative models using accessible R packages. The first is a more flexible semi-Markov model based on semi-parametric Cox regressions for each transition. The second is a simpler, fully-parametric Markov model. Age (at the time of transition) and sex are included as covariates in both models, with different hazard ratios $\exp(\boldsymbol{\beta}_{rs})$ for each $r \to s$ transition. Age-sex interactions were considered and judged not significant.

3.3.1 Semi-parametric, semi-Markov model

In this model,

$$q_{rs}(t, \mathbf{z}(t)) = q^{(0)}_{rs}(t) \exp(\boldsymbol{\beta}'_{rs}\mathbf{z}(t)) \tag{1}$$

thus the hazards are *proportional* between patient groups or covariate values, in other words the covariate value has a constant time-independent multiplicative association with the hazard. If the covariates $\mathbf{z}(t)$ are time dependent, such as age in our example, they are assumed to be step functions which remain constant between each t_j and t_{j+1}.

The baseline hazard $q_{rs}^{(0)}(t)$ is left unspecified and estimated nonparametrically using the Breslow estimator (as in De Wreede *et al.*[50]), and the $\boldsymbol{\beta}_{rs}$ are estimated by maximum partial likelihood. The dependence of $q_{rs}^{(0)}(t)$ on time could be modelled by expressing time t as the time since the start of the process, in this case the date of the first hospital admission. We use an alternative approach of defining t as the time spent by the individual in their current state. Then the function $q_{rs}^{(0)}(t)$ represents how the hazard changes after discharge from hospital, or during a single hospital stay. This is a *clock-reset* or *semi-Markov model* (see Putter *at al.*[14] for further details).

We fit Cox semi-Markov models to the hospital admission data using the `survival` package for R (Therneau and Grambsch[42]). The `mstate` package (De Wreede *et al.*[43]) subsequently computes covariate-specific cumulative hazards for each of the transition-specific Cox models.

3.3.2 Parametric Markov model

In a second, more parsimonious parametric model, the baseline hazard $q_{rs}^{(0)}$ is constant, and the hazard only varies with increasing patient age, included in $\mathbf{z}(t)$. Covariates are again included through proportional hazards.

$$q_{rs}(t, \mathbf{z}(t)) = q_{rs}^{(0)} \exp(\boldsymbol{\beta}'_{rs}\mathbf{z}(t)) \tag{2}$$

Since age is piecewise-constant, the hazard is a step function of time, and the sojourn time in each state r has a piecewise exponential distribution, with a piecewise-constant rate $q_{rr}(t)$. This is a Markov model, since future evolution only depends on the current state. That is, $q_{rs}(t, \mathbf{z}(t), \mathcal{H}_t)$ is independent of $\mathcal{H}_t$.

Again, the $q_{rs}^{(0)}$ and $\boldsymbol{\beta}_{rs}$ are estimated by maximum likelihood, and standard errors are obtained by standard asymptotic theory. Since the state is known to be $S(t_j)$ from t_j until the transition to state $S(t_{j+1})$ at t_{j+1}, the contribution of each patient i to the likelihood is

$$L_i(Q) = \prod_j L_j = \prod_j^{n_i} \exp\{-q_{S(t_j)S(t_j)}(t_{j+1} - t_j)\} q_{S(t_j)S(t_{j+1})} \tag{3}$$

where the q_{rs} in this formula are evaluated for this patient's covariates, assuming their age is constant over this time interval. The complete likelihood $L(Q) = \prod_i L_i(Q)$ is maximised in terms of $\log(q_{rs}^0)$ and $\boldsymbol{\beta}_{rs}$.

This model is fitted using the `msm` package for R (Jackson[44]). This class of models may also be fitted to data where the exact times of transition between states are unknown. This is common in situations where the states are levels of severity of a disease, which may only be known at times of clinic visits (Jackson,[38] Kalbfleisch and Lawless,[44] Kay[18]).

3.4 Prediction from multi-state models

To predict the probability of occupying a particular state at a fixed time in the future, we calculate the transition probability matrix $P(u, t+u)$, where the (r, s) entry of $P(u, t+u)$, $p_{rs}(u, t+u)$, is the probability of being in state s at a time $t+u$, given the state at time u is r. Given $Q(t)$, this is the solution to the Kolmogorov differential equations (see Cox and Miller[45] for further details). For the semi-Markov model, this can be calculated by simulating a large number of individual state histories from the multi-state model given the covariate-specific cumulative hazards, and this can be done by the `mstate` package.[43]

In the parametric model, if the transition intensity matrix Q is constant, given the values of covariates, over the interval $(u, t+u)$, then $P(u, t+u) = P(t)$. In this case, the transition probability matrix can be calculated directly from the matrix exponential of the scaled transition intensity matrix Q scaled by the time interval, i.e., $P(t) = Exp(tQ)$. The transition probability matrix over intervals where Q is piecewise-constant is then calculated as a matrix product of terms like these.

The $P(t)$ can be used to predict the expected total time spent in a state s over a given period of time $(0, T)$, as $E_s = \int_0^T p_{rs}(t)dt$, given that a patient is in state r at time 0. In this study we predict the total time spent in hospital from the first admission until death, a quantity of interest to healthcare providers.

For the parametric model, we can calculate standard errors or confidence intervals for quantities such as these, which are functions of $q_{rs}^{(0)}$ and $\boldsymbol{\beta}_{rs}$, by simulating from the assumed asymptotic normal distribution of the estimators of $q_{rs}^{(0)}$ and $\boldsymbol{\beta}_{rs}$, and recalculating the quantities of interest.[46] Under the semi-Markov model, however, since simulation is required to calculate $P(t)$, a second level of simulation to obtain an accurate confidence interval would be unfeasible.

4 Analysis and Results

4.1 Descriptives

The study cohort consists of $15,298$ patients whose first HF admission ended in 2006. Patients were followed up to December 31st, 2010. Among these individuals, $6,646$ (43.44%) died (from any cause) by the end of the study. The proportion of patients who died during a hospital admission was 8.26%.

Patient age at the time of the first hospitalisation ranged from 18 to 103 years, with mean age (SD) 75.6 (12.6) years. The age of patients at the time of the final discharge ranged from 19 and 105 years, with mean (SD) 76.7 (12.5) years. In the cohort there are $7,184$ (46.96%) males and $8,114$ (53.04%) females. Women were older than men: mean (SD) ages 79.6 (11.4) and 71.5 (12.88), respectively.

The number of admissions to hospital per patient (Table 1) ranged between 1 and 24 (mean = 2.31, median = 2, quantiles 1 and 3). There was no significant difference between men and women in the number of hospitalisations.

	Hospitalizations during follow up					
	1	2	3	4	5	> 5
Number of pts.	15,298	8,891	4,836	2,604	1,492	855
%	100%	58.12%	31.61%	17.02%	9.75%	5.59%

	6	7	8	9	10	> 10
Number of pts.	855	514	302	175	97	56
%	5.59	3.36%	1.97%	1.14%	0.63%	0.37%

Table 1: Distribution of number of admissions to hospital for chronic HF between HF patients and percentage of patients who entered each stage during the 5-year follow up.

Table 2 shows summary statistics for time from the previous discharge to each subsequent admission, for those patients experiencing them. The mean (and median) time to the next hospitalisation decreases as the number of readmissions increases.

	pts.	mean (sd)	median	1Q	3Q	min	max
(1st adm)	15298	—	—	—	—	—	—
to 2nd adm	8891	369.5 (422.4)	180	57.9	558.5	3	1820
to 3rd adm	4836	308.7 (348.81)	160	55	443.3	3	1738
to 4th adm	2604	279.4 (313.13)	154	59	384	4	1691
to 5th adm	1492	238.5 (266.54)	140	50	331	3	1499
to $> 5^{th}$ adm	855	197.5 (216.52)	118	46.5	266	4	1284

Table 2: Summary statistics for times to readmission to hospital for HF-patients.

The overall mean (standard deviation) LOS in hospital is 13.2 (13.9) days (min = 1, median = 9, first and third quantiles respectively equal to 5 and 16, max = 180 days). There is a slight difference between mean LOS of male and female patients (12.9 male vs 13.5 female) and there is no significant difference in LOS among subsequent hospitalisations.

4.2 Multi-state models

The multi-state models described in Section 3 are fitted to the HF data. Table 3 shows the total number of observed transitions for each state. In-hospital mortality increases from 7.16% (first admission in-hospital death rate) up to 9.99% (fifth admission in-hospital death rate), probably due to the aging population and the increasing severity of the HF.

		to $r-th$ discharge	to death
From $r-th$ admission	$r=1$	14,203	1,095
	$r=2$	8,145	746
	$r=3$	4,383	453
	$r=4$	2,378	226
	$r=5$	1,343	149
	$r=5^+$	-	394

		to $(r+1)-th$ admission	to death
From $r-th$ discharge	$r=1$	8,891	1,750
	$r=2$	4,836	980
	$r=3$	2,604	488
	$r=4$	1,492	236
	$r=5$	855	132

Table 3: Transitions for the multi-state model in (2) fitted to HF data.

Associations with age and sex

Figures 2 and 3 show maximum likelihood estimates of the hazard ratios for the effects of age and sex, both for semi-parametric (black) and parametric (red) specifications of the model in (2). An increase of 5 years in age has only a very small effect on readmission and discharge times, decreasing the chance of discharge and increasing the risk of readmission slightly. These estimates are very precise due to the large sample, though are unlikely to hold clinical significance. There is evidence, as expected, that increasing patient age increases the death hazard from all the states. This effect appears to slightly decrease with the number of hospitalisations. This may be due to the fact that as the population ages, it tends to shrink toward more homogeneous and "robust" behaviour ("survival of the fittest"), reducing the apparent contrasts between patients.

In general the gender effect is smaller than the age effect, with few significant hazard ratios. In the earlier stages, women are less likely to change state (die, be admitted or discharged from hospital) than are men. The lower hazard for transitions to death may reflect the longer life-expectancy for women, which the age effect in the model may not have fully adjusted for. These data suggest that there may be a reluctance to admit to hospital women with symptoms of HF in the early stages. Once admitted, women in the early stages of HF were less likely to be discharged early. However once disease severity has reached the later stages, reflected by several admissions, progression through stages and survival is the same for both sexes.

There were some differences between point estimates calculated from the semi-parametric and parametric models although the patterns through the process were similar, and estimates of precision are comparable. The parametric model resulted in an increased effect of age on death out of hospital, and a slightly bigger effect of age on readmission rates, particularly in the early stages of disease. Note the hazards for the parametric model are assumed to be constant within each state. The disagreement between the parametric and semi-parametric models is greatest for the transitions which take place over long periods of time (discharge to readmission or death) for which the hazard may not be constant and there is heavy censoring. The parametric and semi-parametric

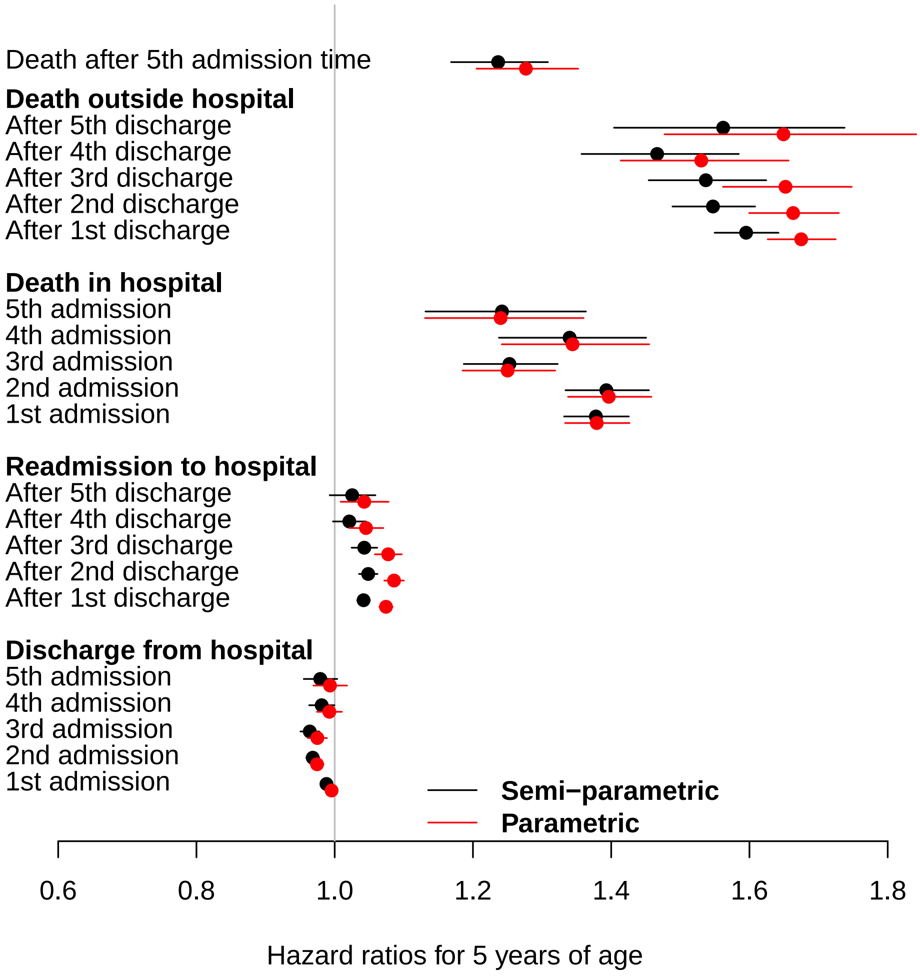


Figure 2: Hazard ratios for a five year increase in age, on each of 21 hospital admission, discharge or death events.

estimates agree for the transitions from admission to discharge or death in hospital, since the hazard is more likely to be constant over the relatively short times spent in hospital, and there is minimal censoring.

Expected survival and time in hospital

Using the methods described in Section 3.4, we estimated the restricted mean survival, and total time spent in any of the five hospital states, over 5 years

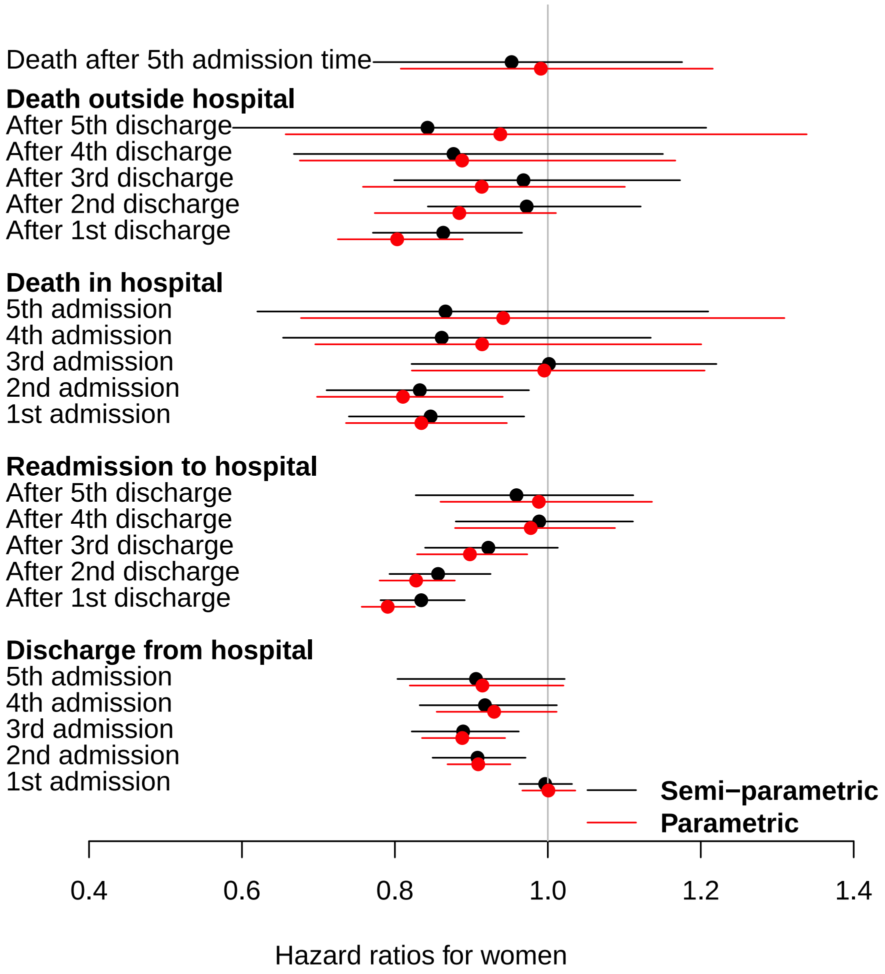


Figure 3: Hazard ratios for a female patient relative to male, on each of 21 hospital admission, discharge or death events.

from the first HF admission. These are shown in Table 4. The mean times spent in hospital for HF treatment, up to the 6th admission, are consistent with just over 2 admissions per patient. A small proportion of patients, 5.6%, will have more than 5 admissions. An advantage of the parametric models is that measures of uncertainty for these estimates are easily available. We expect that the semi-parametric estimates would have similar precision to the parametric estimates, since the confidence intervals for the covariate effects were of similar width (Figures 2 and 3).

	Parametric		Semi-parametric	
	Men	Women	Men	Women
Age	Restricted mean survival (years) over five years			
65	4.32 (4.28, 4.34)	4.44 (4.40, 4.47)	4.33	4.42
70	4.01 (3.97, 4.04)	4.18 (4.14, 4.21)	4.06	4.19
75	3.59 (3.54, 3.63)	3.81 (3.77, 3.85)	3.71	3.84
80	3.06 (3.01, 3.12)	3.33 (3.29, 3.37)	3.24	3.42
85	2.44 (2.38, 2.51)	2.75 (2.70, 2.79)	2.59	2.86
	Expected days spent in hospital over five years			
65	33.74 (33.06, 34.46)	32.36 (31.52, 33.10)	29.97	30.43
70	33.66 (32.98, 34.30)	32.61 (31.92, 33.32)	30.73	29.98
75	32.69 (31.97, 33.36)	32.10 (31.43, 32.74)	30.59	29.82
80	30.63 (29.95, 31.35)	30.60 (29.97, 31.24)	29.01	28.95
85	27.49 (26.69, 28.25)	28.01 (27.37, 28.62)	26.33	26.57

Table 4: Expected survival over five years, and time spent in hospital over five years, by age and sex, under parametric and semi-parametric multi-state models, with 95% confidence intervals where available.

Another advantage of the parametric model is that the mean sojourn times in each state may be estimated. These are the expected times from state entry until transition to another state. Estimates are reported in Table 5 for men and women aged 76 years (the mean population value). Under the semi-parametric model, the hazards are only estimated within the five-year follow-up period of the data, therefore to estimate the mean sojourn times we would need additional parametric assumptions for the hazards beyond that period.

Periods in hospital		
	Male	Female
1st	13.6 (13.4, 14.0)	13.8 (13.5 ,14.1)
2nd	12.3 (12.0, 12.7)	13.7 (13.3 ,14.1)
3rd	12.2 (11.7 ,12.7)	13.6 (13.2 ,14.2)
4th	12.3 (11.6 ,12.8)	13.3 (12.5 ,14.0)
5th	12.8 (12.0 ,13.7)	13.9 (13.2 ,15.0)

Subsequent periods out of hospital		
	Male	Female
1st	676 (662, 696)	820 (802, 842)
2nd	568 (548, 589)	657 (639, 681)
3rd	508 (483, 522)	555 (531, 574)
4th	419 (396, 452)	434 (399, 462)
5th	340 (317, 360)	346 (313, 369)

Table 5: Estimated mean sojourn times, in days, for each transient state of the parametric multi-state model, with 95% confidence intervals. Age on state entry is set to the mean population value (76 years).

This table shows that mean stay in hospital does not change substantially as the number of admissions increases. However the times between admissions do decrease, reflecting an acceleration in the disease process once it has been diagnosed and has resulted in an initial admission. Mean sojourn times for women were slightly longer, consistent with the hazard ratios which showed that women were less likely to change states than men. We could hypothesise that women are more likely to have carer commitments at home and so may only be admitted for more severe HF episodes, resulting in slightly longer stay. This and other hypotheses could be examined in future clinical studies.

4.2.1 Model assessment

Since every transition time is known, we can calculate Kaplan-Meier estimates of the time from state entry until the next transition for particular age-sex subgroups. Estimates from the fitted models for the corresponding covariate category can be compared with these to assess model fit (as discussed by Titman and Sharples.[16] For patients in hospital, both the semi-parametric and parametric models give good predictions of the probability of remaining in hospital for a 76 year old patient (Figure 4). This is consistent with the agreement of the corresponding covariate effects between the models in Figure 2. Figure 5 shows that the semi-parametric model accounts better for the decrease in the hazard of readmission (or death) since the time of last discharge. This is because the parametric assumptions for the hazards are likely to be reasonable over the short times spent in hospital, but not over longer periods. The parametric model assumes the hazards vary only with age, whereas the semi-parametric model relaxes the Markov assumption by also modelling the hazards as non-parametric functions of the time spent in the current state.

For out-of-hospital starting states, the extent of censoring and the sparsity of data at later times increases with the number of admissions, therefore any visible discrepancy between the Kaplan-Meier and fitted curves at these times is less likely to be significant. Any remaining lack of fit of the semi-parametric model may result from non-proportional hazards. A test of the correlation of the Schoenfeld residuals with the Kaplan-Meier estimates at the corresponding time showed that hazards were only significantly non-proportional for two out of the 19 transition-specific Cox models, and then only for the age effect. Since the sizes of these effects (discharge after 1st admission, and readmission after 1st discharge in Figure 2) are not clinically significant, this is not a concern.

5 Conclusions

Contemporary administrative health care databases allow for a new kind of epidemiological research, based on real-time availability and low-cost data. Despite the issues surrounding the reliability of such data, in the last decade significant improvements have been obtained in this area, and the use of administrative databases in clinical biostatistics has become an accepted practice. The benefits of using these data for health system planning and evaluation go far beyond the fact that they are cheap and quickly available: they are population based, comprehensive, capture real health system use, longitudinal, and can be linked to other data. Even if it can be difficult to properly define a population of interest starting from these databanks, administrative databases represent a valuable clinical resource. At the same time, they represent a great challenge for statistics and statistical models.

In this work we focused on the use of administrative data for gaining insights into the impact of heart failure. We used multi-state models to simultaneously predict survival, time to the next hospitalisation and total time spent in hospital, and how these depend on age, gender and hospitalisation history.

For a chronic disease, such as CHF, in-hospital states are heavily controlled by the health care provider and assumptions of constant hazards and proportional hazards for these states are likely to be valid. However the out of hospital state is determined by a range of influences including the underlying progression of the disease, comorbidities and the ageing of the population. These factors are not adequately modelled using a parametric model based on constant hazards, although the bias in estimates of the hazard ratios was not large in this population. Thus if the focus is on estimation of covariate effects, constant hazards may be an adequate approximation, but for studies that focus on assessment of time to readmission, and associated health care consumption, it is not a reliable approach.

We were able to show that times between hospital admissions decreased as the number of admissions increased, reflecting HF progression, and to quantify expected times between admissions. As might be expected, patients who were older at first admission were readmitted more frequently, as were men (compared with women) in the earlier stages of HF. However, the number of admissions and associated time spent in hospital, over this 5 year period was roughly constant with age, decreasing only for age of onset of around 85 years. For example, as a proportion of the restricted mean survival time over 5 years, time in hospital ranged from about 1.9% for 65 year old patients to 2.5% for 85 year old women.

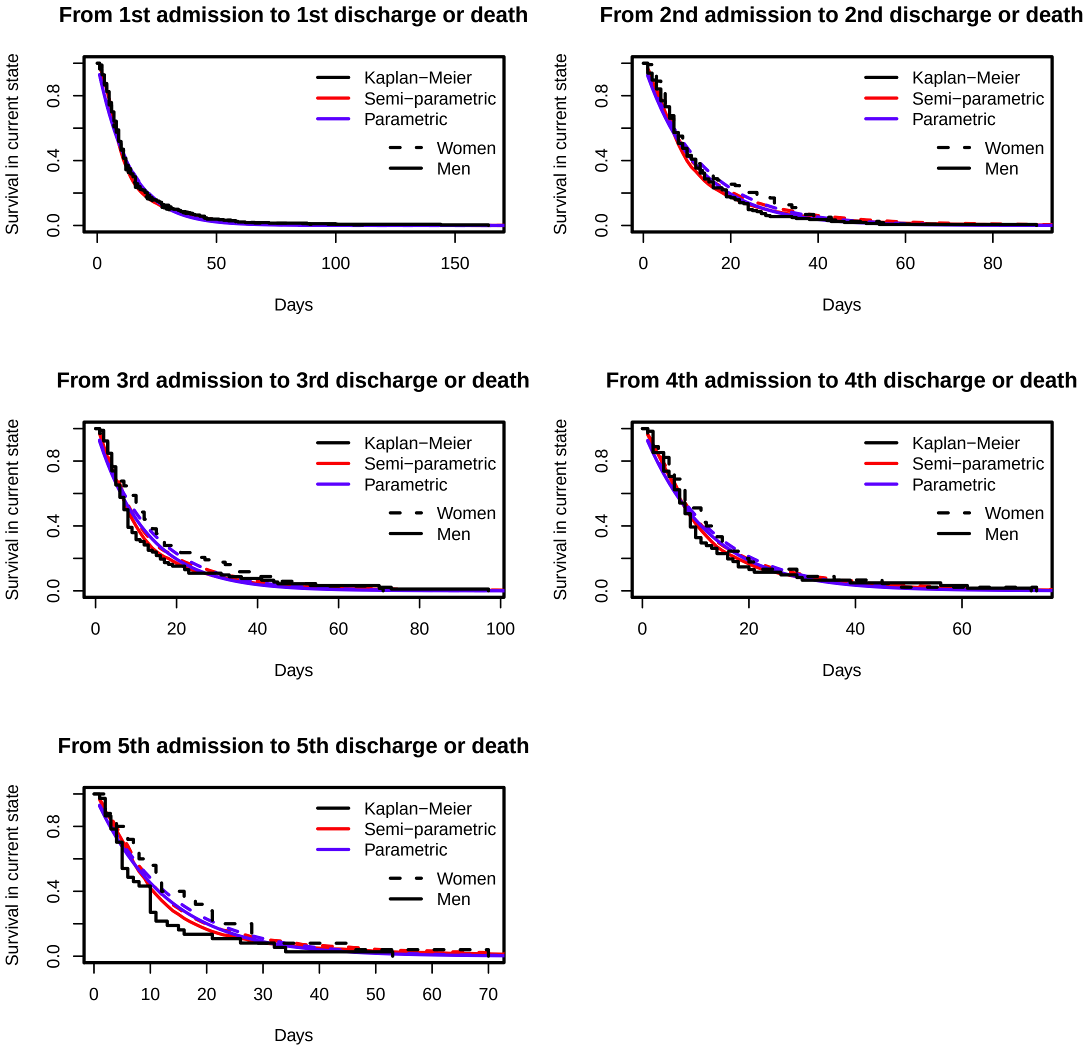
Due to the size of the Lombardia administrative databases it is possible to study a range of factors influencing health care consumption, through jointly modelling hospital admissions and death. This has resulted in precise estimates of expected survival times, times spent in hospital and covariate effects. Additionally there is sufficient power to investigate interactions between covariates, which has not been possible with smaller data registries. In this study there were no significant interaction effects between age and sex on model parameters, and the size of the dataset ensures that we can be confident in this assertion.

Multi-state models are effective in describing clinical processes as discrete states. Nevertheless, due to the difficulty in inference for some types of data, strong assumptions on the process dynamics and on covariate effects are often applied. As pointed out in Titman and Sharples,[16] it is difficult to make universally valid recommendations on model checking as often the model assumptions depend on the particular application. For example, the Markov assumption claims that given the present state, the future evolution of the process (hospitalisations as a proxy for HF progression, in the case of interest) is independent of the states previously visited and the transition times among them. This assumption is often restrictive and when it fails the model may provide inconsistent estimates. In this work we checked this assumption through informal diagnostic plots. Rodriguez-Girondo and De Una-Alvarez,[47] proposed a formal test for the Markov assumption in the illness-death model, based on measuring the future-past association over time through generalisations of Kendall's τ, but no solutions are present, to the best of our knowledge, for more general multi-state models.

Despite these restrictive assumptions, Markov models are often a convenient starting point for jointly modelling hospital admissions and death. The `msm` package in `R`, among others, has made implementation straightforward for a wide range of model structures, and in particular for intermittently-observed multi-state data where the exact times of transition are unknown. In addition, estimates of covariate effects were only slightly biased in our application. For situations such as the fitting of serial hospitalisation, in which transitions between states are fully observed, the `mstate` package implements more flexible semi-parametric and/or semi-Markov models. These models provide less biased estimates of sojourn times and covariate effects, but require computationally-expensive simulations from the fitted model to provide estimates of quantities of interest. Another advantage of parametric models is to estimate quantities that require extrapolation beyond the time horizon of the data, such as (unrestricted) mean survival, or mean sojourn times in our example.

Acknowledgments

This work is within the Project of Ricerca Finalizzata "Utilization of Regional Health Service databases for evaluating epidemiology, short- and medium-term outcome, and process indexes in patients hospitalized for heart failure", funded by the Italian Ministry of Health and Regione Lombardia - Healthcare division.



From 1st admission to 1st discharge or death

From 2nd admission to 2nd discharge or death

From 3rd admission to 3rd discharge or death

From 4th admission to 4th discharge or death

From 5th admission to 5th discharge or death

Figure 4: Kaplan-Meier (black) curves of time to discharge or death, from each in-hospital starting state, and estimated probabilities of remaining in that state from parametric (red) and semi-parametric (blue) models, for women and men aged 76 years at day 0.

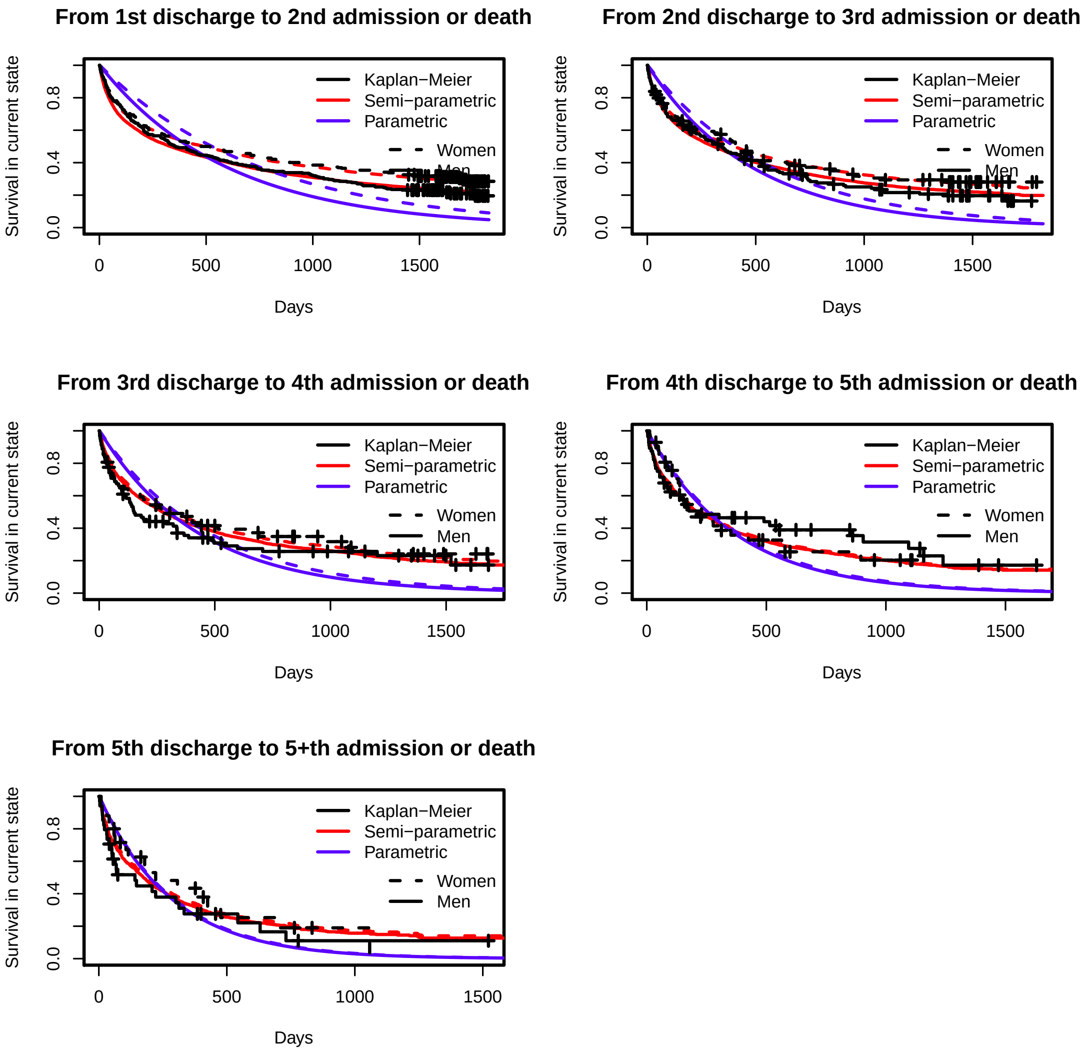


Figure 5: Kaplan-Meier (black) curves of time to readmission or death, from each out-of-hospital starting state, and estimated probabilities of remaining in that state from parametric (red) and semi-parametric (blue) models, for women and men aged 76 years at day 0.

References

[1] Ho KK, Pinsky JL, Kannel WB, Levy D. The epidemiology of heart failure: the Framingham Studyl. *Journal of the American College of Cardiology*. 1993;22(4 Suppl A):6A.

[2] Cowie MR, Mosterd A, Wood DA, et al. The epidemiology of heart failure. *European Heart Journal*. 1997;18:208–215.

[3] Mosterd A, Deckers JW, Hoes AW, Nederpel A, Smeets A, Linker DT, et al. Classification of heart failure in population based research: an assessment of six heart failure scores. *European Journal of Epidemiology*. 1997;13(5):491.

[4] Roger VL. The heart failure epidemic. *International Journal of Environ Res Public Health*. 2010;7(4):1807–1830.

[5] Bleumink GS, Knetsch AM, Sturkenboom MC, Straus SM, Hofman A, Deckers JW, et al. Quantifying the heart failure epidemic: prevalence, incidence rate, lifetime risk and prognosis of heart failure The Rotterdam Study. *European Heart Journal*. 2004;25(18):1614.

[6] McMurray JJ, Petrie MC, Murdoch DR, Davie AP. Clinical epidemiology of heart failure: public and private health burden. *European Heart Journal*. 1998;19(Suppl):9.

[7] Lloyd-Jones D, Adams RJ, Brown TM, Carnethon M, Dai S, De Simone G, et al. ;.

[8] ISTAT - Istituto Nazionale di Statistica [homepage on the Internet]; [cited 2014 Jun 21]. Available from: `http://demo.istat.it/`.

[9] Castaneda J, Bart G. Appraisal of several methods to model time to multiple events per subjecys: modelling time to hospitalizations and death. *Revista Colombiana de Estadistica*. 2010;33(1):43–61.

[10] Andersen PK, Keiding N. Multistate models for event history analysis. *Statistical Methods in Medical Research*. 2002;11:91–115.

[11] Hougaard P. Multi-state Models: a Review. *Lifetime Data Analysis*. 1999;5:239–264.

[12] Commenges D. Inference for multistate models from interval-censored data. *Statistical Methods in Medical Research*. 2002;11:167–182.

[13] Cook RJ. A mixed model for markov processes under panel observation. *Biometrics*. 1999;55:178–183.

[14] Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multistate models. *Statistics in Medicine*. 2007;26:2389–2430.

[15] Sommmen C, Alioum A, Commenges D. A multistate approach for estimating the incidence of human immunodeficiency virus by using HIV and AIDS French surveillance data. *Statistics in Medicine*. 2009;28:1554–1568.

[16] Titman AC, Sharples LD. Model diagnostics for multi-state models. *Statistical Methods in Medical Research*. 2009;19:621–651.

[17] Duffy SW, Chen HH. Estimation of mean sojourn time in breast cancer screening using a Markov chain model of entry to and exit from preclinical detectable phase. *Statistics in Medicine*. 1995;14:1531–1543.

[18] Kay R. A Markov Model for Analysing Cancer Markers and Disease States in Survival Studies. *Biometrics*. 1986;42:855–865.

[19] Chen B, Yi GY, Cook RJ. Analysis of interval-censored disease progression data via multi-state models under a nonignorable inspection process. *Statistics in Medicine*. 2010;29(11):1175–1189.

[20] Commenges D, Joly P. Multi-state model for dementia, institutionalization and death. *Communications in Statistics - A*. 2004;33:1315–1326.

[21] Sutradhar R, Forbes S, Urbach DR, Paszat L, Rabeneck L, Baxter NN. Multistate models for comparing trends in hospitalizations among young adult survivors of colorectal cancer and matched controls. *BMC Health Service Research*. 2012;12:353.

[22] Innovative Care for Chronic Conditions: Building Blocks for Action. Global Report [homepage on the Internet]; 2002 [cited 2014 Jun 21]. World Health Organization, Geneva, Switzerland. Available from: `www.who.int/diabetesactiononline/about/icccglobalreport.pdf`.

[23] Postmus D, Van Veldhuisen DJ, Jaarsma T, Luttik ML, Lassus J, Mebazaa A, et al. The COACH risk engine: a multistate model for predicting survival and hospitalization in patients with heart failure. *European Journal of Heart Failure*. 2012;14(2):168–175.

[24] Barbieri P, Grieco N, Ieva F, Paganoni AM, Secchi P. In: Exploitation, integration and statistical analysis of Public Health Database and STEMI archive in Lombardia Region. Complex data modelling and computationally intensive statistical methods. Series - Contribution to Statistics. Springer; 2010. p. 41–56.

[25] Wirehn AB, Karlsson HM, Cartensen JM, et al. Estimating Disease Prevalence using a population-based administrative healthcare database. *Scandinavian Journal of Public Health*. 2007;35:424–431.

[26] Saczynski JS, Andrade SE, Harrold LR, Tjia J, Cutrona SL, Dodd KS, et al. A systematic review of validated methods for identifying heart failure using administrative data. *Pharmacoepidemiology and drug safety*. 2012;21(S1):129–140.

[27] Macchia A, Monte S, Romero M, D'Ettorre A, Tognon G. The prognostic influence of chronic obstructive pulmonary disease in patients hospitalised for chronic heart failure. *European Journal of Heart Failure*. 2007;9:942–948.

[28] Au AG, McAlister FA, Bakal JA, Ezekowitz J, Kaul P, van Walraven C. Predicting the risk of unplanned readmission ordeath within 30 days of discharge after a heart failure hospitalization. *American Heart Journal*. 2012;164(3):365–372.

[29] Aylin P, Bottle A, Majeed A. Use of administrative data or clinical databases as predictors of risk of death in hospital: comparison of models. *BMJ*. 2007;334:1044.

[30] Philbin EF DT. Prediction of Hospital Readmission for Heart Failure: Development of a Simple Risk Score Based on Administrative Data. *Journal of the American College of Cardiology*. 1999;33(6).

[31] Lee Douglas S, Donovan L, Austin PC, Yanyan G, Liu PP, Rouleau JL, et al. Comparison of coding of heart failure and comorbidities in administrative and clinical data for use in outcomes research. *Medical Care*. 2005;43(2):182–188.

[32] Quach S, Blais C, Quan H. Administrative data have high variation in validity for recording heart failure. *Canadian Journal of Cardiology*. 2010;26(8).

[33] Schultz SE, Rothwell DM, Chen Z, Tu K. Identifying cases of congestive heart failure from administrative data: a validation study using primary care patient records. *Chronic Diseases and Injuries in Canada*. 2013;13(3).

[34] Muggah E, Graves E, Bennett C, Manuel DG. Ascertainment of chronic diseases using population health data: a comparison of health administrative data and patient self-report. *BMC Public Health*. 2013;13.

[35] Iron K, Lu H, Manuel D, Henry D, Gershon A. Using Linked Health Administrative Data to Assess the Clinical and Healthcare System Impact of Chronic Diseases in Ontario. *Healthcare Quarterly*. 2011;14(3):23–27.

[36] R Development Core Team. R: A Language and Environment for Statistical Computing; 2009. Available from: `http://www.R-project.org`.

[37] Therneau TM. A Package for Survival Analysis in S. *R package version 237-7*. 2014;.

[38] Jackson CH. Multi-State Models for Panel Data: The msm Package for R. *Journal of Statistical Software*. 2011;38(8):1–29.

[39] AHRQ QualityIndicators. Guide to InpatientQualityIndicators: Quality of Care in Hospitals - Volume, Mortality, and Utilization. . 2007;Version 3.1.

[40] Pope GC, Kautter J, Ellis RP, AshJohn AS, Ayanian Z, Iezzoni LI, et al. Risk Adjustment of Medicare Capitation Payments Using the CMS-HCC Model. *Health Care Financial Review*. 2004;25(4):119–141.

[41] Pope GC, Kautter J, Ingber MJ, Freeman S. Evaluation of the CMS-HCC RiskAdjustment Model - Final Report. *RTI International for CMS*. 201;.

[42] Therneau TM, Grambsch PM;.

[43] De Wreede LC, Fiocco M, Putter H. mstate: an R package for the analysis of competing risks and multi-state models. *Journal of Statistical Software*. 2011;38(7):1–30.

[44] Kalbfleisch J, Lawless JF. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*. 1985;80(392):863–871.

[45] Cox DR, Miller HD. *The Theory of Stochastic Processes*. London: Chapman and Hall; 1965.

[46] Mandel M. Simulation-Based Confidence Intervals for Functions With Complicated Derivatives. *The American Statistician*. 2013;67(2):76–81.

[47] Rodriguez-Girondo M, De Una-Alvarez J. A nonparametric test for Markovianity in the illness-death model. *Statistics in Medicine*. 2012;31:4416–4427.

MOX Technical Reports, last issues

Dipartimento di Matematica "F. Brioschi",
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

24/2014 Ieva, F., Jackson, C.H., Sharples, L.D.
Multi-State modelling of repeated hospitalisation and death in patients with Heart Failure: the use of large administrative databases in clinical epidemiology

23/2014 Ieva, F., Paganoni, A.M., Tarabelloni, N.
Covariance Based Unsupervised Classification in Functional Data Analysis

22/2014 Arioli, G.
Insegnare Matematica con Mathematica

21/2014 Artina, M.; Fornasier, M.; Micheletti, S.; Perotto, S.
The benefits of anisotropic mesh adaptation for brittle fractures under plane-strain conditions

20/2014 Artina, M.; Fornasier, M.; Micheletti, S.; Perotto, S.
Anisotropic mesh adaptation for crack detection in brittle materials

19/2014 L.Bonaventura; R. Ferretti
Semi-Lagrangian methods for parabolic problems in divergence form

18/2014 Tumolo, G.; Bonaventura, L.
An accurate and efficient numerical framework for adaptive numerical weather prediction

17/2014 Discacciati, M.; Gervasio, P.; Quarteroni, A.
Interface Control Domain Decomposition (ICDD) Method for Stokes-Darcy coupling

15/2014 Esfandiar, B.; Porta, G.; Perotto, S.; Guadagnini, A;
Anisotropic mesh and time step adaptivity for solute transport modeling in porous media

16/2014 Dede, L.; Jaggli, C.; Quarteroni, A.
Isogeometric numerical dispersion analysis for elastic wave propagation