Monitoring a developing pandemic with available data

María Luz Gámiz Pérez
Department of Statistics and Operations Research, University of Granada, Spain

Enno Mammen
Institute of Applied Mathematics, Heidelberg University, Germany

María Dolores Martínez Miranda*
Department of Statistics and Operations Research, University of Granada, Spain

Jens Perch Nielsen
Cass Business School, City, University of London, UK

October 21, 2021

Abstract

When a pandemic is developing then data collection is chaotic and the most important data one week might come from a completely different place the next week. Therefore, it is important that both input and output are easy to understand. This paper provides a dynamic approach to monitoring the most important transitions in a pandemic. The data used are simple reflecting the type of data almost every news reading person learned to know during the Covid-19 pandemic. The simplicity of the data is a challenge and new missing data methodology has to be developed. The methodology is illustrated via the case of Covid-19 developing in France.

Keywords: Hazard; Marker ;Missing data; Local linear estimation; DO-validation

*Corresponding author: mmiranda@ugr.es

1 Introduction

Balancing complexity is at heart of mathematical statistics. When plenty of data are available it is possible to work with complicated mathematical statistical models; reduced complexity is necessary when data are sparse. This trade-off has many names, but it is often referred to as balancing signal to noise. There is another aspect of balancing complexity and simplicity that has been given less attention in mathematical statistics. In this aspect it is the complexity of the data gathered that is considered rather than the complexity of the statistical model itself. When trying to monitor a developing pandemic, data collection time, communication and knowledge sharing across communities, quick data processing and easily understood output are all extremely important. A global pandemic is likely to develop in a chaotic way and where information and knowledge have to be collected within different environments on an almost daily basis. One day the outbreak of the pandemic might be in China, the next week in Italy and then moving to all other countries from there. In other words: both input and output of the methodology used should be standardized and easily exchanged. The new methodology of this paper is introducing a dynamic extension in various directions of the recent paper Gamiz *et al.* (2022). These generalizations enable us to monitor a developing pandemic with available data. Gamiz *et al.* (2022) provides a new technique solving a new missing link data problem for survival analysis and uses it on Covid-19 pandemic data. The missing link is between information of arrivals and information on leavers in a dynamic system. Available Covid-19 pandemic data counted arrivals and leavers, but were missing the link between the arrivals and the leavers, so when someone left the hospital for example, information was not registered on time spent in hospital before leaving. This paper works with the same type of data and therefore has a similar missing data problem. However, it turns out that the statistical methodology derived in Gamiz *et al.* (2022) almost immediately extends to a dynamic setting considered in this paper.

2 The Dynamic Missing Link Survival Model and the Dynamic Missing Inhomogenous Poisson Model

In this Section a formal definition of the dynamic mathematical statistical model is given. In Section 2.1 and Section 2.2 the formal model is described in the mathematical tractable situation where continuous data are available. It is easy to follow the problem of missing information in this setting.

In Section 2.1 the framework is such that any arrival gives rise to a random number of future events and there might not be a direct link between the arrivals and the future events. The arrivals almost play the role as a covariate or as exposure and every arrival gives rise to one independent inhomogenous Poisson process of future events. The model in Section 2.1 is useful when modelling the stochastic nature of infections.

Section 2.2 is only about the situation where there is connected exactly one leave to one arrival. We call this the survival model setting. This model formulation is useful when considering duration of time in hospital during a pandemic for example.

In Section 2.3 and Section 2.4 we provide a discrete approach that is very similar to the approach in Section 2.1 and 2.2 but that can be used when only daily data are available as in our available Covid-19 pandemic data.

2.1 The Dynamic Missing Link Inhomogeneous Poisson Model

Let us assume that subjects (from a population of size $\mathcal{N}$) arrive to a system at random times modelled by a counting process $\tilde{N}_1$. Specifically $\tilde{N}_1(t)$ counts the number of subjects that enter the system during the interval $(0, t]$ and associated to the ith subject a stochastic process $\{Z_{1,i}(t), t \geq 0\}$ is defined, where $Z_{1,i}(t)$ takes value 1 when the subject enters the system at any time in the interval $(0, t]$, and 0 otherwise. In Section 8 this process will be considered as a covariate or marker process. Then $\tilde{N}_1(t) = \sum_{i=1}^{\mathcal{N}} Z_{1,i}(t)$.

Each subject entering the system gives rise to a new counting process $N_{1,i}$ that can take values in the set $\{0, 1, 2, \ldots\}$. This counting process starts at the time of arrival and has a jump of size 1 each time the event under study is happening. The event of interest is considered a recurrent event in the sense more than one occurrence can be registered

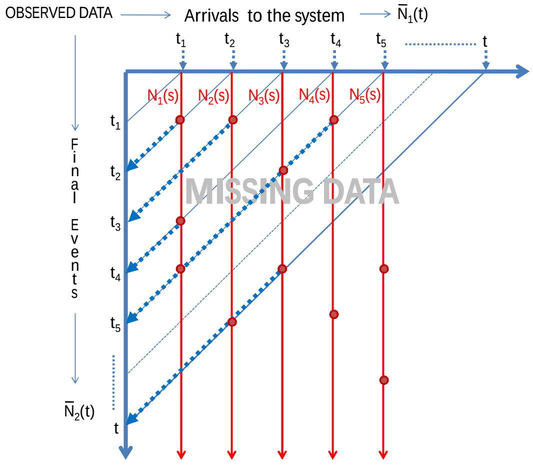

Figure 1: The missing data problem with inhomogeneous Poisson modelling.

(for example a repairable failure) associated to a particular subject that enters the system. Then $N_{1,i}$ is assumed to be an inhomogeneous Poisson process. The situation is as in Figure 1.

Figure 1 gives a simple graphical description. The available data are the counting processes $\tilde{N}_1$ and $\tilde{N}_2$. The arrivals are represented on the horizontal axis Figure 1 by the counting process $\tilde{N}_1(t)$. More than one event of interest can be recorded related to each subject entering the system, then the total counts of final events registered by calendar time are represented on the vertical axis of the plot by the counting process $\tilde{N}_2(t)$.

We assume that the intensity function of the process $N_{1,i}$ can be written $\lambda_{1,i}(s) = \alpha_1(s, Z_{1,i}(s))$, where $\alpha_1(\cdot,\cdot)$ is an unknown (deterministic) hazard function. When full information is available of the stochastic processes $Z_{1,i}$ and $N_{1,i}$, one could estimate the hazard function α_1, by the marker dependent estimator that will be defined later in Section 3.1. However we do not observe the stochastic process $N_{1,i}$ directly. Instead, we observe the counting process $\tilde{N}_1 = \sum_{i=1}^{\mathcal{N}} Z_{1,i}$ above counting arrivals and the counting process $\tilde{N}_2(t)$ defined as follows

$$\tilde{N}_2(t) = \sum_{\{i:Z_{1,i}(t)=1\}} N_{1,i}(t - t_i)$$

where $t_i = \inf\{t : Z_{1,i}(t) = 1\}$, for the i-th subject.

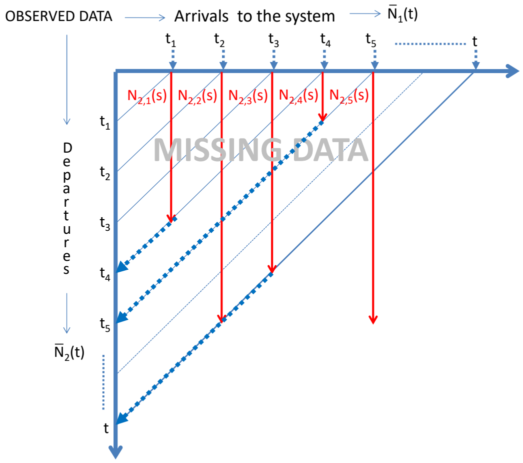

Figure 2: Missing data problem with *non-recurrent* events.

2.2 The Dynamic Missing Link Survival Model

Similarly to the situation in Section 2.1, let us assume that subjects (from a population of size $\mathcal{N}$) arrive to a system at random times modelled by a counting process $\bar{N}_1$. That is, $\bar{N}_1(t)$ counts the number of subjects that enter the system during the interval $(0, t]$. Again, associated to the ith subject we define a stochastic process $\{Z_{2,i}(t), t \geq 0\}$, where $Z_{2,i}(t)$ takes value 1 when the subject enters the system at any time in the interval $(0, t]$, and 0 otherwise. We introduce a different notation for the covariate process to highlight that the arrivals of this type originate events inside the system which are of a very different nature compared to the ones described in the previous section. Then $\bar{N}_1(t) = \sum_{i=1}^{\mathcal{N}} Z_{2,i}(t)$.

Each subject entering the system gives rise to a new counting process $N_{2,i}$ that can only take values in $\{0, 1\}$. This counting process starts at the time of arrival jumping to one when the event under study is happening (if this event is happening at all). After the occurrence of the event of interest, the subject is supposed not to be at risk any further and then no more occurrences associated to this particular subject are registered (for example, a non-repairable failure). The situation is as in Figure 2. In this case, a survival model is considered. $N_{2,i}$ is the counting process associated to the survival time of the i-th subject, and the corresponding risk process $Y_{2,i}$ taking value 1 when the subject is at risk and 0 otherwise, is defined.

Figure 2 is a graphical simplification of the real situation. The available data are two counting processes $\bar{N}_1$ and $\bar{N}_2$. The arrivals are represented on the horizontal axis of Figure 2 by the counting process $\bar{N}_1(t)$. There is at most one occurrence of the final event for each subject entering the system, then we interpret it as a departure from the system. The total counts of departures by calendar time are represented on the vertical axis by the counting process $\bar{N}_2(t)$.

We assume that the intensity function of the process $N_{2,i}$ can be written $\lambda_{2,i}(s) = \alpha_2(s, Z_{2,i}(s))Y_{2,i}(s)$, where $Y_{2,i}$ is a predictable process taking value 1 when the subject i is at risk (inside the system) and 0 otherwise, and $\alpha_2(\cdot, \cdot)$ is an unknown (deterministic) hazard function.

When full information is available of the stochastic processes $Z_{2,i}$, $N_{2,i}$ and $Y_{2,i}$, one could estimate the hazard function α_2, by the marker dependent estimator given in Nielsen (1998) and detailed in Section 3.2. We do not observe the stochastic processes $N_{2,i}$ and $Y_{2,i}$ directly. Instead, we observe the counting process $\bar{N}_1 = \sum_{i=1}^{\mathcal{N}} Z_{2,i}$ above counting arrivals and the counting process $\bar{N}_2(t)$ defined as follows

$$\bar{N}_2(t) = \sum_{\{i:Z_{2,i}(t)=1\}} N_{2,i}(t - t_i)$$

where $t_i = \inf\{t : Z_{2,i}(t) = 1\}$, for the i-th subject.

2.3 The Dynamic Missing Link Inhomogeneous Poisson Model with discrete data

Often occurrences and exposures are not observed continuously and the only data available is discretely collected during pre-specified time intervals. In the following we introduce discretized observed versions of the above continuous counting processes: $\tilde{N}_1(t)$, for the arrivals of subjects to the system; and, $\tilde{N}_2(t)$, for the number of occurrences of the event of interest until time t, when observed, not continuously, but on a discrete grid of time points. This grid not necessarily has to be equidistant, but for simplicity in the notation and without any loss of generality, we consider the set $\{1, 2, \ldots, M\}$.

Let us define, for $x = 1, 2, \ldots, M$,

$$E_{x,1} = \int_{x-1}^{x} d\tilde{N}_1(u);$$

the total number of subjects entering the system in the interval $(x-1, x]$; and,

$$O_x = \int_{x-1}^{x} d\tilde{N}_2(u);$$

are the total number of occurrences registered in the system in the interval $(x-1, x]$.
As mentioned above, each subject that enters the system originates a Poisson process associated to its arrival in the system, i.e. $N_{1,i}$ for the ith subject arriving at time t_i. Let $N_{1,x}$ be the aggregated counting process for all subjects arriving in the interval $(x-1, x]$, that is

$$N_{1,x}(s) = \sum_{\{i:t_i \in (x-1,x]\}} N_{1,i}(s)$$

for $s > 0$. When the processes $N_{1,i}$ are observed (full information), we can also define

$$O_{x,d} = \int_{d-1}^{d} dN_{1,x-d+1}(s)$$

for $1 \leq d \leq x \leq M$, which are the total number of events registered at time x from all individual Poisson processes that started at time $x - d + 1$.
It is satisfied that $O_x = \sum_{d=1}^{x} O_{x,d}$, for all $1 \leq x \leq M$.
When full information is available, these counts are of course directly observable. However, our missing link data problem implies that full data information is not available, and we are left to do our analysis with occurrence counts data like O_x above together with discrete exposure data like $E_x = \sum_{r=1}^{x} E_{r,1}$, for $x = 1, 2, \ldots, M$. The size of the exposure at time x with a duration in the system exactly equal to d is $E_{x,d} = E_{x-d+1,1}$ for $1 \leq d \leq x \leq M$, and it can be checked that that $E_x = \sum_{d=1}^{x} E_{x,d}$.

2.4 The Dynamic Missing Link Survival Model with discrete data

In this section we introduce notation for discretized observed versions of the above continuous counting processes: $\bar{N}_1(t)$, for the arrivals of subjects to the system; and, $\bar{N}_2(t)$, for the number of occurrences of the event of interest, when observed, not continuously, but on a discrete grid of time points. Again this grid not necessarily has to be equidistant, but for simplicity, we consider the set $\{1, 2, \dots, M\}$.

Let us define, for $x = 1, 2, \dots, M$,

$$E_{x,1} = \int_{x-1}^{x} d\bar{N}_1(u);$$

the total number of subjects entering the system in the interval $(x-1, x]$; and,

$$O_x = \int_{x-1}^{x} d\bar{N}_2(u);$$

are the total number of departures from the system registered in the interval $(x-1, x]$. Each subject that enters the system originates a counting process associated with its survival time inside the system, $N_{2,i}$ for the ith subject arriving at time t_i. Let $N_{2,x}$ be the aggregated counting process for all subjects arriving in the interval $(x-1, x]$, that is

$$N_{2,x}(s) = \sum_{\{i: t_i \in (x-1,x]\}} N_{2,i}(s)$$

for $s > 0$. When the processes $N_{2,i}$ are observed (full information), we can also define the following counts

$$O_{x,d} = \int_{d-1}^{d} dN_{2,x-d+1}(s)$$

for $1 \leq d \leq x \leq M$, which are the total number of events registered at time x from all individual survival counting processes that started at time $x - d + 1$.

As can be seen in Gamiz *et al.* (2022) we have that $O_x = \sum_{d=1}^{x} O_{x,d}$, for all $1 \leq x \leq M$. When full information is available, these counts are of course directly observable. However, our missing link data problem implies that full data information is not available, and we are left to do our analysis with occurrence counts data like O_x above together with discrete exposure data like E_x below.

$$E_x = \sum_{r=1}^{x} E_{r,1} - \sum_{r=1}^{x-1} O_r$$

for $x = 1, 2, \dots, M$.

The number of subjects still at risk at time x and that have stayed in the system for a time exactly equal to d is

$$E_{x,d} = E_{x-d+1,1} - \sum_{s=1}^{d-1} O_{x-s,d-s}$$

for $1 \leq d \leq x \leq M$. We have that $E_x = \sum_{d=1}^{x} E_{x,d}$.

3 Estimation of the two-dimensional intensities for transition function when full information is available

In this section we provide the two-dimensional intensity estimators we will use when we have full information. The estimator was defined in Nielsen (1998) for the survival model and we consider an adaptation for the intensity of the inhomogeneous Poisson process.

Let $\{(N_1, Z_1, Y_1), \ldots, (N_n, Z_n, Y_n)\}$ be independent and identically distributed processes, where N_i is counting process with respect to an increasing, right continuous, complete filtration $\mathcal{F}_t$, $t \in (0, \tau)$; Z_i is a covariate process; and, Y_i is a predictable risk process. The random intensity of the process N_i is modelled as

$$\lambda_i(t) = \alpha(Z_i(t), t)Y_i(t), \tag{1}$$

with $t \in (0, \tau)$ and no restriction on the functional form of $\alpha(\cdot, \cdot)$.

For simplicity we denote $W_i(s) = (Z_i(s), s)$. Let $\mathcal{K}$ be a two-dimensional kernel and $\bar{b} = (b_1, b_2)$ a bandwidth vector. Let $\mathcal{K}_{\bar{b}}(x-y) = K_{1,b_1}(x_1-y_1)K_{2,b_2}(x_2-y_2)$, where $x = (x_1, x_2)$ and $y = (y_1, y_2)$ and $K_{j,b_j}(\cdot) = K_j(\cdot)/b_j$, with K_j a general univariate kernel, $j = 1, 2$.

Let us define the following moments. A scalar function

$$\mathrm{A}_0(x) = \sum\nolimits_{i=1}^{n} \int_0^\tau \mathcal{K}_{\bar{b}}\{x - W_i(s)\}Y_i(s)ds; \tag{2}$$

$\mathrm{A}_1(x)$ is a vector function whose jth component is

$$A_{1,j}(x) = \sum\nolimits_{i=1}^{n} \int_0^\tau \mathcal{K}_{\bar{b}}\{x - W_i(s)\}\left(x_j - W_{ij}(s)\right)Y_i(s)ds, \tag{3}$$

for $j = 1, 2$; and a matrix function $\mathbf{A}_2(x)$ with dimension 2×2 whose (j, k)-element is given by

$$A_{2,jk}(x) = \sum\nolimits_{i=1}^{n} \int_0^\tau \mathcal{K}_{\bar{b}}\{x - W_i(s)\}\left(x_j - W_{ij}(s)\right)\left(x_k - W_{ik}(s)\right)Y_i(s)ds, \tag{4}$$

for $j, k = 1, 2$.

3.1 The two-dimensional intensity estimator of the inhomogeneous Poisson model

In this subsection, we define the adjustment of the two-dimensional local linear intensity estimator of Nielsen (1998) to the inhomogeneous Poisson case we need for the transitions defined in Subsection 8.1 and Subsection 8.2 below.

Let $N_1, \ldots, N_n$ be n independent inhomogeneous Poisson processes. For each individual i, the process N_i has intensity $\lambda_i(t) = \alpha(Z_i(t), t)$, which is a special case of expression (1) for $Y_i = 1$, for all $i = 1, 2, \ldots, n$. The moments defined in (2)-(4) simplify in this particular case and the local linear estimator of α given in Nielsen (1998) can be adapted and written as

$$\widehat{\alpha}_{K,\bar{b}}(x) = \sum_{i=1}^{n} \int_0^{\tau} \frac{\{1 - u^t \mathbf{A}_2(x)^{-1} \mathrm{A}_1(x)\}}{\mathrm{A}_0(x) - \mathrm{A}_1(x)^t \mathbf{A}_2(x)^{-1} \mathrm{A}_1(x)} \mathcal{K}_{\bar{b}}\{x - W_i(s)\} dN_i(s). \qquad (5)$$

3.2 The two-dimensional hazard estimator of the survival model

In this subsection, we describe the original local linear marker dependent hazard estimator of Nielsen (1998) that we need for the transitions defined in Subsections 8.1, 8.2 and 8.3 below.

In this case, $Y_i(t)$ is a random variable taking value 1 when the subject i is at risk and under observation at time t, and 0 otherwise. We only take moments $\mathrm{A}_1(x)$ and $\mathbf{A}_2(x)$, defined in (3)-(4). The local linear estimator of α is given by

$$\widehat{\alpha}_{K,\bar{b}}(x) = \frac{\sum_{i=1}^n \int_0^\tau \{1 - u^t \mathbf{A}_2(x)^{-1} \mathrm{A}_1(x)\} K_{x,\bar{b}}\{x - W_i(s)\} dN_i(s)}{\sum_{i=1}^n \int_0^\tau \{1 - u^t \mathbf{A}_2(x)^{-1} \mathrm{A}_1(x)\} K_{x,\bar{b}}\{x - W_i(s)\} Y_i(s) ds}. \qquad (6)$$

3.3 The two-dimensional intensity estimator of the inhomogeneous Poisson model with discrete data

(To be written)[1]

[1] Should we write the entire discrete data approach in an Appendix?

3.4 The two-dimensional hazard estimator of the survival model with discrete data

(To be written)

4 Generating exposure and occurrences with available data from an initial guess

4.1 Generating occurrences from an initial guess of the inhomogeneous Poisson process

Following the notation in Section 2.1, given that the subject i arrives in the system at time z_i, an inhomogeneous Poisson process starts $N_{1,i}$ with intensity rate $\lambda_{1,i}(s) = \alpha(z_i, s)$, for $i = 1, \dots, n$.[2] Since no observations of the processes $(N_{1,i}, Z_{1,i})$ are available we cannot directly estimate the intensity $\alpha(\cdot, \cdot)$. In this section we present a procedure to generate this necessary information to obtain the two-dimensional estimator, defined in Section 3.1, given that we observe the process $\widetilde{N}_1(t)$ counting the new arrivals in the system in the interval $(0, t]$, as well as the process $\widetilde{N}_2(t)$ counting all the events that occur inside the system in the interval $(0, t]$. To do it we start with a prior guess about the unknown intensity, $\alpha_0(\cdot, \cdot)$, which we take as if it were the true rate of the process $N_{1,i}$.

Let us denote ε_1 the intensity rate of new arrivals and $\mathcal{F}_t$ the σ-algebra containing all the history of events occurring in the system in the interval $(0, t]$. Then

$$E\left[d\widetilde{N}_2(t)|\mathcal{F}_t\right] = \left(\int_0^t \alpha_0(t-s, s)d\widetilde{N}_1(t-s)\right)\, dt,$$

or, noticing that $d\widetilde{N}_1(t-s) \approx \varepsilon_1(t-s)ds$,

$$E\left[d\widetilde{N}_2(t)|\mathcal{F}_t\right] = \left(\int_0^t \alpha_0(t-s, s)\varepsilon_1(t-s)ds\right)\, dt.$$

[2] In Section 3.1 we write the intensity as $\alpha(Z_i(s), s)$, that is, depending on $Z_i(s)$ which is a 0-1 process that indicates the time at which the i-th subject arrives in the system. Now we use the time $z_i = \min\{s : Z_i(s) = 1\}$ as the first argument of α instead of $Z_i(s)$.

After some calculations, we can write, for $0 < s < t$ arbitrary though fixed,

$$E\left[d\widetilde{N}_2(t)|\mathcal{F}_t\right] = \frac{\alpha_0(t-s,s)\varepsilon_1(t-s)dt}{\frac{\alpha(t-s,s)\varepsilon_1(t-s,s)}{\left(\int_0^t \alpha_0(t-s,s)\varepsilon_1(t-s)ds\right)}}.$$

We define the following function

$$q(t,s) = \frac{\alpha_0(t-s,s)\varepsilon_1(t-s)}{\left(\int_0^t \alpha_0(t-u,u)\varepsilon_1(t-u)du\right)}$$

and then

$$E\left[d\widetilde{N}_2(t)|\mathcal{F}_t\right] q(t,s) = \alpha_0(t-s,s)\varepsilon_1(t-s,s)dt. \tag{7}$$

As we know, each arrival generates a Poisson process so that the size of the exposure at time t due to subjects arriving at as small interval around $t-s$ only depends on the total number of arrivals at that time, that is $d\widetilde{N}_1(t-s) \approx \varepsilon_1(t-s)ds$. Then the instantaneous probability of an event occurring inside the system at time t which is associated to subjects entering at time $t-s$ is given by

$$E\left[dN_{1,t-s}(s)|\mathcal{F}_t\right] = \alpha_0(t-s,s)\varepsilon_1(t-s)ds$$

This expression together with (7) motivates the following

$$E\left[dN_{1,t-s}(s)|\mathcal{F}_t\right]dt = E\left[d\widetilde{N}_2(t)|\mathcal{F}_t\right] q(t,s)ds. \tag{8}$$

Finally, for a subject i entering the system at a small interval around time $t-s$ we denote $N_{1,i}(s) = N_{1,t-s}(s)$. Then we can approximate $dN_{1,i}(s) \approx E\left[dN_{1,t-s}(s)|\mathcal{F}_t\right]$, which, based on equation (8), can be obtained from available data thus providing the information needed to get an estimator of the intensity $\widehat{\alpha}_{K,\bar{b}}$ as described in (5).

4.2 Generating exposure and occurrences from an initial guess of the survival model

Following the definitions in Section 2.2, let us denote $(Z_{2,i}, N_{2,i}, Y_{2,i})$ the model associated to the survival time inside the system spent by the i-th subject. We assume that $N_{2,i}$ is a counting process with intensity $\lambda_{2,i}(s) = \alpha(z_i,s)Y_{2,i}(s)$. Since no observations of this model are available we cannot directly estimate the hazard $\alpha(\cdot,\cdot)$. We consider a prior candidate for the hazard function, i.e. α_0, which is treated as the true model. In this section we present a procedure to generate the information necessary to obtain the two-dimensional hazard estimator given in Section 3.2, from the observation of processes $\bar{N}_1(t) = \sum_{i=1}^{\mathcal{N}} Z_{2,i}(t)$ counting the arrivals in the system; and $\bar{N}_2(t) = \sum_{\{i:Z_{2,i}(t)=1\}} N_{2,i}(t-z_i)$, with $z_i = \inf\{t : Z_{2,i}(t) = 1\}$. Let us denote ε_1 the intensity of $\bar{N}_1$, and $\mathcal{F}_t$ is the σ-algebra that contains all the history of the two processes until time t.

Again we have $d\bar{N}_1(t-s) \approx \varepsilon_1(t-s)ds$ and then

$$E\left[d\bar{N}_2(t)|\mathcal{F}_t\right] = \left(\int_0^t \alpha_0(t-s,s)S_0(t-s,s)\varepsilon_1(t-s)ds\right)\, dt, \tag{9}$$

with $S_0(z,s) = \exp\left\{-\int_0^s \alpha_0(z,u)du\right\}$, for all $z, s > 0$.

After some manipulation in equation (9) we get

$$E\left[d\bar{N}_2(t)|\mathcal{F}_t\right] q(t,s) = \alpha_0(t-s,s)S_0(t-s,s)\varepsilon_1(t-s)dt, \tag{10}$$

where, we denote

$$q(t,s) = \frac{\alpha_0(t-s,s)S_0(t-s,s)\varepsilon_1(t-s)}{\int_0^t \alpha_0(t-u,u)S_0(t-u,u)\varepsilon_1(t-u)du}.$$

The expected number of subjects remaining in the system at time t among those who entered at a small interval around $t-s$ can be calculated as $E\left[Y_{t-s,2}(s)\right] = S_0(t-s,s)\varepsilon_1(t-s)$, then the instantaneous probability of a subject leaving the system immediately after t can be approximated as $E\left[dN_{2,t-s}(s)|\mathcal{F}_t\right] \approx \alpha_0(t-s,s)S_0(t-s,s)\varepsilon_1(t-s)ds$. Then, from equation (10) we obtain

$$E\left[dN_{2,t-s}(s)|\mathcal{F}_t\right] dt = E\left[d\bar{N}_2(t)|\mathcal{F}_t\right] q(t,s)ds. \tag{11}$$

The number of subjects at risk at time t, regardless how long they have remained in the system, is $\bar{Y}_2(t) = \sum_{\{i:z_i\le t\}} Y_{2,i}(t-z_i)$ and the total expected exposure at time t can be calculated

$$E\left[\bar{Y}_2(t)\right] = \int_0^t S_0(t-s,s)\varepsilon_1(t-s)ds. \tag{12}$$

Some transformations in this equation lead to

$$E\left[\bar{Y}_2(t)\right] h(t,s) = S_0(t-s,s)\varepsilon_1(t-s). \tag{13}$$

where

$$h(t,s) = \frac{S_0(t-s,s)\varepsilon_1(t-s)}{\int_0^t S_0(t-u,u)\varepsilon_1(t-u)du},$$

and then

$$E\left[Y_{2,t-s}(s)\right] = E\left[\bar{Y}_2(t)\right] h(t,s). \tag{14}$$

Finally, for a subject i entering the system at a time in a small interval around $t-s$ we denote $N_{2,i}(s) = N_{2,t-s}(s)$ and $Y_{2,i}(s) = Y_{2,t-s}(s)$. Then we can approximate $dN_{2,i}(s) \approx E\left[dN_{2,t-s}(s)|\mathcal{F}_t\right]$ and $Y_{2,i}(s) \approx E\left[Y_{2,t-s}(s)\right]$. Using equations (11)and (14) we can generate this information from available data which allows us to obtain an estimator of the hazard $\widehat{\alpha}_{K,\bar{b}}$ as described in (6).

4.3 Generating occurrences from an initial guess of the inhomogeneous Poisson process with discrete data

(To be written)

4.4 Generating exposure and occurrences from an initial guess of the survival model with discrete data

(To be written)

5 General considerations when monitoring and forecasting in a dynamic environment

Our dynamic modelling is broadly speaking based on two types of transitions. One type where the number of individuals are well defined and a follow-up type survival analysis is possible. And then another type of transition where the number of individuals involved are biased by dynamic definitions and underestimation. We use two different mathematical approaches to describe these two types of transitions.

5.1 A transition where the number of individuals involved are well known

In a confusing pandemic where definitions and measurements and other things are changes by the hour, there are stable components. It is for example often well defined what it means to be hospitalized even though the criteria for hospitalized might develop dynamically. And the daily number of hospitalized people can be expected to be recorded quite well in many countries. Death is also well defined of course even though death-by-the-pandemic can have a dynamic definition changing with a developing dynamic. We there decide to model transitions from individual transitions from number of hospitalized to death-in-hospital or recovery-from-hospital via something that looks like standard survival analysis.

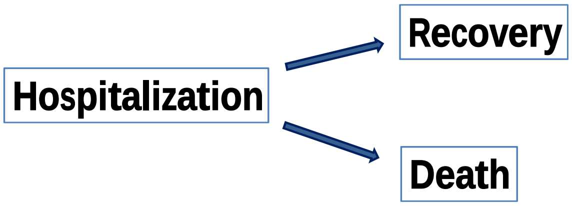

Figure 3: Transition from hospital to death or recovery.

Technically speaking we cannot use standard survival analysis on the above transition when dealing with what we call "available data". Our available data contain information of number of hospitalized every day, numbers of deaths-in-hospital every day and number of recovery-in-hospital every day. But available data does not contain information on the link between these events. We therefore do not know directly from data which hospital-durations died or recovered hospitalized had on one particular day of measurement. However, this statistical problem can be overcome by introducing some new survival techniques on missing data introduced by this paper and discussed in the methodological section.

5.2 Another type of transition where the number of individuals involved are biased by dynamic definitions

There is another type of transition where the number of individuals involved are based on some subsample of the population and even with a dynamic criteria for selecting this subsample. Such a transition is the transition from number of infected individuals to number of hospitalized.

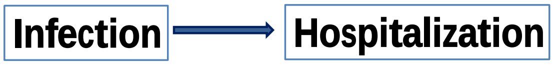

Figure 4: Transition into the hospitalized state.

We model such a system in another way via an inhomogeneous poison process interpreting the number of infected as a dynamic indicator or covariate rather than as a number of individuals.

5.3 Constructing a statistical transition system using the two above type of transitions

When defining a statistical transition system using the above two transition options, then the exercise is to use the first type of transition as much as possible and also to try to use transitions that are short and easy to estimate. In other words, it might be better to estimate a transition from infected to hospitalized and combine it with a transition from hospitalized to death, than to go directly from number of infected to death.

5.4 Principles of forecasting in a dynamic environment

Forecasting is of course always tricky. Let us define and indicator called C_t indicating at any point in time t whether the future will be different or equal to the immediate past.

If C_t equals one then we can forecast the immediate future based on the immediate past. However, in a pandemic there are many change-points of severity. There are long periods where little is being done (and C_t might have a tendency to increase slowly) and there might be few but very important change-points where there measures (e.g. lock down) are introduced to minimize future infections (and C_t might drop dramatically in a matter of days). The point of our approach to forecasting in this paper is that we can monitor and forecast the pandemic when C_t equals one. In that situation a dynamic statistical methodology as introduced in this paper can do the job via a surprisingly simple data collection reflecting what can be considered "available data" in many countries. Therefore, the only thing left to monitor and forecast well is to have a dynamic point of view and the constant C_t. That work needs expert advice specific to countries or local regions within countries. This important work cannot be dismissed via a statistical analysis based on simplified available data. But it is important that the experts involved should only consider forecasting and understand the C_t rather than being responsible for a full statistical model.

6 Estimating with available data. The case of France

In this section we present our main ideas on monitoring a developing pandemic based on available data. We use the recent Covid-19 pandemic and the country France as our case study. We walk through our modelling principles and the new mathematical statistical inventions necessary to implement our new approach. However, we do defer all mathematical definitions and theorems and proofs till later sections and the appendix.

6.1 Time in hospital

As mentioned in Section 5.3 it is important to construct the statistical modelling of the developing pandemic such that robust components get as much weight as possible.

Time spent in hospital is such a robust component and we have decided to start our analysis here and build the rest of the dynamic system around this important central component. When analysing *time spent in hospital*, the exact number of individuals entering and leaving hospitals are assumed to be observed every day together with the additional information on how many die at hospital every day. Notice that the collected data is aggregated in nature and individual follow-up data is not collected. We do not assume to observe which exact individuals who are leaving or recovering at one particular day. This implies that standard survival methodology does not apply and we have had to introduce a new missing data methodology to the tool-box of survival analysis. The mathematical and methodological details of this new survival analysis technique is deferred to Section A.1 in the Appendix.

The most easy way to follow our idea is to look at an example. Consider the transition from being-in-hospital to dying-in-hospital or recovering-from-hospital displayed in Figure 3 of Section 1.1.

6.1.1 Examples of concrete hospital transitions, the Covid-19 case of France

Time spent in hospital is a dynamic concept in the sense that it depends on the particular date an individual is admitted to hospital. Figure 5 displays the estimations of the hazard function of time in hospital until recovery (left panel) or death (right panel) for individuals entering at different dates. We have used a two-dimensional hazard estimator. One dimension being the date of admission and the other the duration-time-in-hospital. For example, the solid black line in the left panel represents the hazard function of the duration until recovery for patients who enter the hospital on 30-September. The hazard function can be interpreted as the probability a patient leaves the hospital due to recovery conditioned on the duration of his/her stay. The solid black line shows a decreasing tendency as time in hospital for these patients passes. In concrete, we see that the about 6.5% of patients that enter the hospital on the 30-September receive clinical discharge on the same day. Also, we can say that after a stay of 10 days, a person who was admitted to hospital on the 30-September has a probability about 0.06 of recovery, while the probability of recovery was barely of 0.035 on the 10-May for patients admitted on the 30-April and with a stay of 10 days (see dashed red line). Finally, for people entering the hospital on the 31-July, the probability of recovering at any date later does not depend on the length of the stay. That is, a person who enter the hospital on the 31-July and still remains in hospital on the 5-August has probability of leaving with clinical discharge of about 0.046; if he/she still remains in hospital by the 20 of August again has a probability of leaving with clinical discharge of about 0.046. That is, for patients entering the hospital on the 31-July, the time until recovery shows a constant hazard rate, at least during the first 35 days of stay.

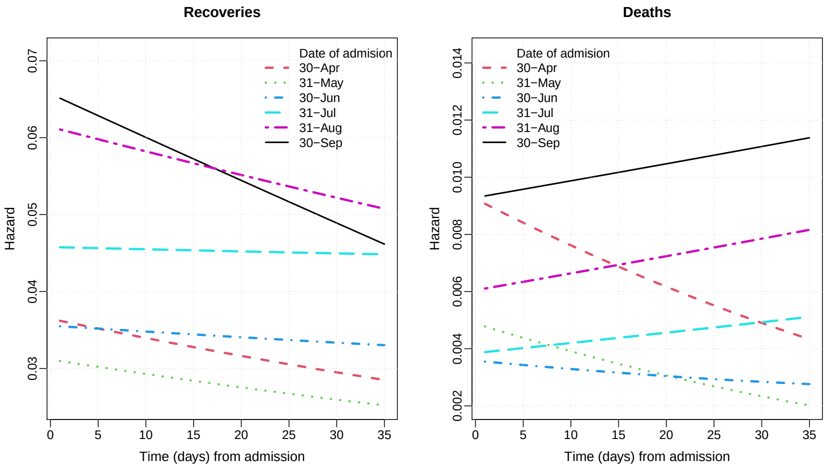

Figure 5: Hazard rate of time spent in hospital for individuals entering the hospital at different dates in the period from March to November in 2020. Left panel: Hazard rate for time since admission until hospital discharge. Right panel: Hazard rate of time since admission until death.

The first victim caused by SARS-Cov-2 in Europe was officially reported in France on 24 of January 2020. On 11 of March 2020 the WHO declared COVID-19 as a world pandemic and from that time on many countries started to introduce contention measures aimed at controlling the spread of the virus. In particular on 16 March 2020, French authorities implemented strict measures such as closing schools, universities and non-essential services as well as mobility restrictions and isolation. The efficiency of the measures was proven by a significant descent of the number of cases in the following weeks with the consequent relief of the national sanitary system directly implying a significant improvement of the healthcare. This can be seen on the right panel of Figure 5, where the hazard mortality rate for new hospitalized is represented for different admission date. The graph in Figure 5 (right panel) shows that for people just arriving on the 30 of April the risk of death is below 1% and decreases as time in hospital passes. This risk is half this figure for people arriving to the hospital one month later. This can be an indicator of the collapse of the hospitals in the months immediately after the pandemic outbreak (March and April). It is changing dynamically what it means to be a hospitalized patient during a pandemic. Political decisions, available hospital resources, fatigue of hospital employees, and many other things play a dynamic role together with the changing character of the way the pandemic itself develop in the population. Therefore, any definition in this paper is meant to be time dependent. What it means to be hospitalized or recover or even dying-as-infected might be different for two different calendar times. Our mathematical formulation given below fully accommodate such dynamics. The calendar time dependency of our two-dimensional marker dependent hazard is the key tool to achieve our fully flexible dynamic modelling of a developing pandemic. When it comes to our specific example of Covid-19 in France it seems evident that there is an improvement in the clinical experience and a perhaps also a better understanding of the disease from the first wave of the pandemic in the spring till the second wave of the pandemic in the fall.

We can see this looking at the left panel of Figure 5. From May to October, the conditional probability for a person to receive hospital discharge, given he/she has been in hospital for d days, shows an increasing tendency as a function of admission date, for any value of d. For example, for patients just arrived, $d = 1$, the probability of leaving hospital due to recovery ranges from 3% on the 31-May to almost 7% on the 30-September.

The average daily number of new hospitalizations was around four times as high in May 2020 compared to October 2020. This dynamics implied a significant pressure on hospitals in October changing the overall chances of recovering from the infection while in hospital. When studying recoveries and deaths while in hospital in Figure 5 and the age-dependent version in Figure 6, it seems very clear that the probability of getting out of hospital alive is changing with the dynamics of the pandemic. We therefore conclude that any modelling of a developing pandemic has to be able to work with dynamic definitions of the stages involved in the pandemic as well as with dynamically developing underlying statistical parameters as we do in this paper.

There are different dynamics depending on age groups. Figure 6 shows the variations in instantaneous calendar time dependent probability of death given duration in hospital (the time dependent hazard rate) for patients across different age groups.

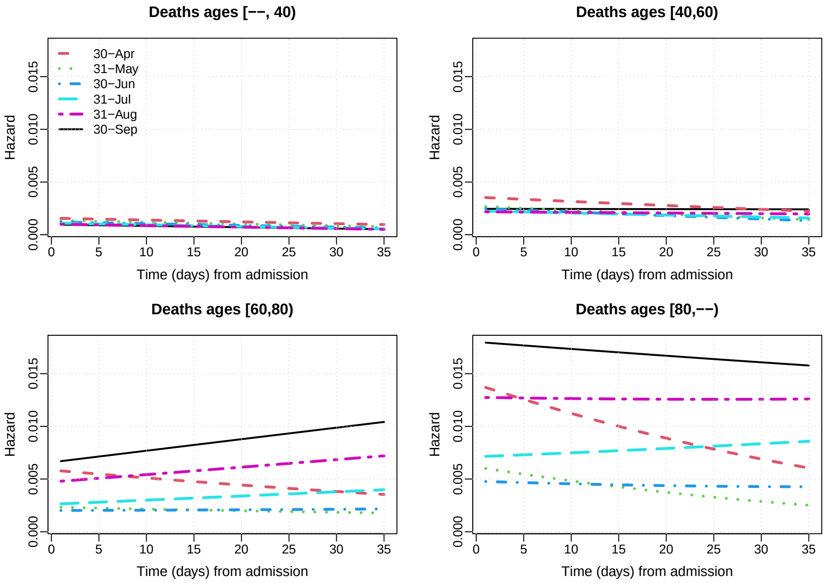

Figure 6: Hazard rate of time spent in hospital for individuals entering the hospital at different dates in the period from March to November in 2020. Left panel: Hazard rate for time since admission until hospital discharge. Right panel: Hazard rate of time since admission until death

The above hazard rate estimators provide us with sufficient information to calculate the expected time in hospital for a given individual patient admitted to hospital at some given date. See Figure 7 where the information is disaggregated by age groups. A prominent peak for people hospitalized by the end of May (shown by all age groups) highlights the fact of the changing behaviour of covid on distinct times of the calendar since the pandemic outbreak. As can be seen, age is an important covariate, and, among other things, while the different conditions under which the pandemic has evolved (variants of the virus and different restrictions regimes) do not have apparently serious impact on the younger population, it seems to strongly affect people in the oldest groups for which the date of infection is key.

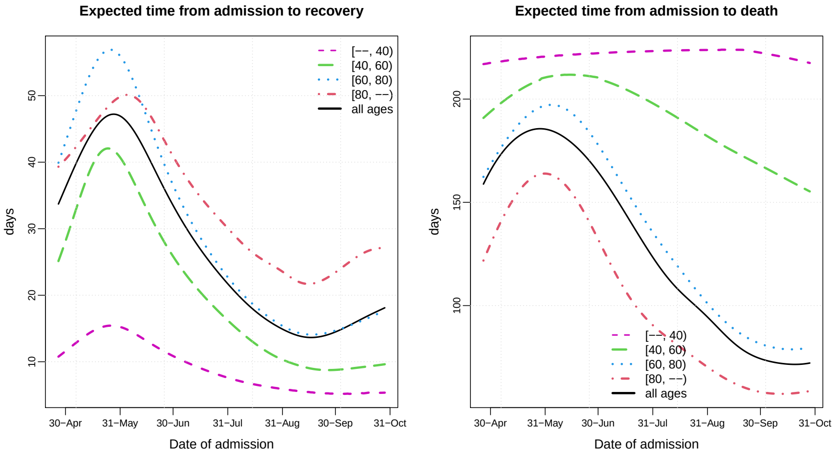

Figure 7: Expected time in hospital by admission date from March to November in 2020. Left panel: Expected time since admission until hospital discharge. Right panel: Expected time since admission until death

Variations over time of durations of stays in hospital until recovery or death are partly explained by changes in the age of patients. In short, it can be said that the length of stays in hospital for recoveries has decreased from the first wave in about five days for the full sample. The reduction is bigger for people between 60 and 80 years and less significant for ages below 40. The length of stays for deaths has also decreased from about 150 days during the first wave until less than 100 days in the second wave, for the full sample, being the reduction is less significant for the younger patients.

Figure 8 helps answer the following question: What is the probability that a subject who has been in hospital for d days can leave it alive?

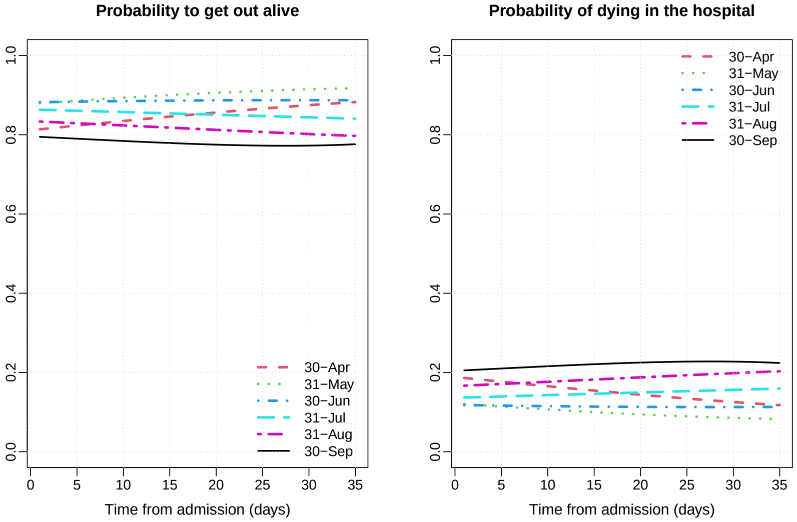

Figure 8: Probability of outcome by cause specific from March to November in 2020. Left panel: Probability of leaving the hospital due to recovery. Right panel: Probability of leaving the hospital due to death

The probability of getting out alive slightly increases with hospitalization time. This increase is more substantial for people above 80 years in the first month in hospital. The probability of dying in hospital decreases with hospitalization time. In the first day in hospital almost 40% of the older people (above 80) will die, this percentage is about 20% for people between 60 and 80 years, and below 10% for the younger.

6.2 From infected to being in hospital

The next transition we consider is the transition from being infected to entering hospital. We do not know the number of infected people and we do not have follow up data of individuals infected (see Figure 4). What we do have is a vague indicator of infected individuals. An indicator that is changing over time. This indicator is the number of individuals tested positive in the pandemic. Clearly this indicator is higher when there is more testing and lower when there is little testing activity. And also the composition of individuals being tested might change over time, where for example, as we saw in the Covid-19 pandemic, elderly people get tested relatively much in the beginning where testing in general is low, while then younger people get tested much more later in the pandemic where testing is more frequent and young people might need a negative test to socialize in the weekend. In other words, we cannot make a direct connection between our infection indicator and the number of infected, we only have a dynamic indicator that is changing in its nature over time. But we can assume that this change is smooth and gradual over time and we can therefore do something by introducing a dynamic smooth system that is changing over time with the changing indicator.

6.2.1 Examples of concrete transitions from infected to hospital, the Covid-19 case of France

Because of the vague nature of the definition of the infection-indicator we have to use transition methodology exposed in Section 1.2 when investigating the transition from being infected to entering hospital. While the input is a vague indicator of infections in the population, the output is more direct, namely number of people entering hospital.

After the first months of the pandemic, it was estimated that 70% of infected individuals manifested the disease, and for around a 30% of these, the illness progression was so serious that they needed hospitalization. This means that the hospitalization rate was roughly estimated by a 0.21, considering all the patients until that moment and regardless the exact time they were infected (citation). As far as the pandemic has lasted over time several mutations of the virus have arisen causing different implications in patients. For example, some mutations may spread more easily or show signs of resistance to existing treatment. Conditions change and it is natural to understand the number of hospitalized per unit time as a dynamic concept. The more spread of the virus the more infected people and, in principle, the more people needing to go to hospital. But the mean age of cases has decreased in the second wave compared to the period March-April, probably due to the massive screening rolled out later, so incrementing the number of detected patients asymptomatic or with mild symptoms who do not need special healthcare. In fact, we have seen from data that during the month of October the hospitalized supposed only 4.4% of all positive tested during this month while it was 24.5% in June, even when the number of cases has increased a tenfold in October.

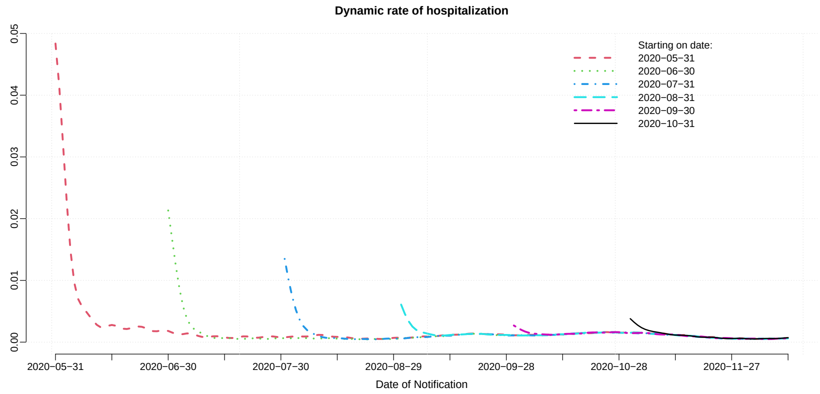

Figure 9: Dynamic estimation of the rate of hospitalization.

Figure 9 illustrates the use of our methodology to estimate the rate of hospitalization in a dynamic context. We introduce below a new local linear two-dimensional estimator for the intensity rate of an inhomogeneous Poisson process. While the local linear marker dependent hazard estimator has been known for a while, see Nielsen(1999), then it seems that our paper is the first to introduce the local linear marker dependent hazard estimator for an inhomogenous Poisson process, see Section 4 for more details. We now assume that, starting from a particular day z the future number of hospitalized arrive following an inhomogeneous Poisson process whose intensity rate depends on the number of infected detected on the day z and use our new local linear estimator to quantify and visualize the dynamics of the infection of the French Covid-19 pandemic. For example, the red dashed line is the rate of hospitalizations starting on the 31 of May while the black solid line is the corresponding to those starting on the 31 of September. At first sight, Figure 9 shows a perhaps surprising issue: a sharp descending trend of the hospitalization rates when are seen as a function of the date of onset. That is, almost 5% of new cases is estimated to be hospitalized on the 31 of May; on the 30 of June, there will be around 2%; and, this figure is below 0.5% on the 31 of September. This might suggest a slowdown in the speed of arrivals to hospital with time which may be strange given the dramatic increase in number of cases in the last period, but can be explained due to a loss of capacity of the hospitals as the pandemic continues with time and patients accumulate so that less of them enter hospital and also because the variation of the virus that circulated during the second wave was spreading faster but was not more lethal. It is important to note that the age of patients was significantly lower in the second wave.

6.3 Examples of concrete transitions from infected to infected, the Covid-19 case of France

When it comes to the relationship between infections in the population at one point in time and then infections in the population at a later point in time, then it is perhaps surprising that the methodology described in Section 1.2 can be used once again. Now both the input and the output are vague indicators of number of infections.

Figure 10: Transition from Infected to Infected.

The number of observed infected in a developing pandemic vary significantly in response to, among other things, the implementation of unprecedented interventions (lockdowns, social distance, etc.). Besides, the same number of observed infected might have completely different interpretation in the beginning of the pandemic and after say a year of the pandemic, when testing has become more available and is perhaps also more accurate. The number of observed infected is of course only a fraction of the total number of infected, so it underestimate the spread of the pandemic, especially at the time of the outbreak. But the observed number of infected is a time-dependent indicator of how serious the pandemic is at the moment. This is the reason our model is using number of infected as a fundamental measure of the size of an inhomogeneous Poisson process starting at this time. We use again a two-dimensional marker dependent rate to capture the dynamics of the infection process. More specifically, new positives reported on a particular day give rise to new infected in the future. Then, every day of the calendar a stochastic process that counts the new infected from that day on is originated, so that we have a family of Poisson processes. Each process of he family is indexed (marked) by the date at which it starts and the corresponding intensity function is obtained using a two-dimensional estimator.

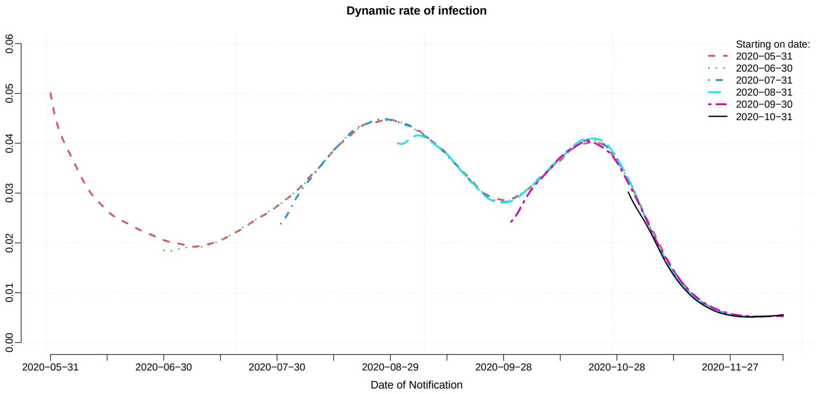

Figure 11: Dynamic estimation of the rate of infection.

Figure 11 is a demonstration of how the virus has spread from May to November. The intensities of the processes that start at some particular dates are represented. It is noticeable the steep decline in the number of cases in both May and November as a result of the severe measures imposed in France at those moments in time.

6.4 Relationship between dying in hospital and dying outside hospital , the Covid-19 case of France

Based on the above transitions, we are now able to understand transitions from infected to infected, and from infected to hospital and from hospital to recovery or death. Even though we are only using available data. However, we are still not able to say anything about the total number of deaths in the population due to the pandemic. To do this we need as available data the additional information on the daily number of people dying outside the hospital due the pandemic. We then operate with a dynamic ratio between number of people dying in hospital at a given date and number of people dying outside hospital at a given data.

6.4.1 Examples of smooth developments of number of deaths in hospital, the Covid-19 case of France

The curve in Figure 12 is a smooth estimation of the density of the number of the deaths reported by some hospital in France. The curve has been obtained using a local linear kernel density estimator that additionally considers a bias correction. It can be seen that the days with highest number of deaths was reported during the first wave, in early April. However, the second wave that far more affected younger populations, caused higher mortality among these age groups, that is, the deaths were more frequent inside the hospitals than in residences for the oldest. With the implementation of vaccination plans, the curve showed a clear descent later in February.

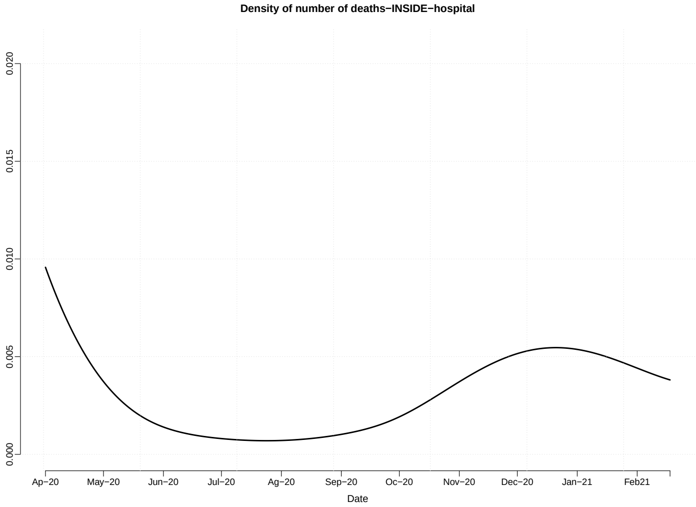

Figure 12: Smooth density estimation of the number of deaths-inside-hospital from April-2020 to February-2021.

6.4.2 Examples of smooth developments of number of deaths outside hospital, the Covid-19 case of France

The curve in Figure 13 is a smooth estimation of the density of number of reported deaths occurred in social medical establishments for the elderly (denominated *EHPAD* and *EMS* in France), that is, deaths that are not registered in hospitals. Again we have used a local linear kernel density estimator with a bias correction to obtain the curve. As can be deduced from the curve, the number of deaths occurred was very high at the beginning of the pandemic, in the month of April, then these establishments were put under the surveillance and very strict lockdowns were implemented in these centres until the situation was under control. Although a slight rise in deaths can be seen at the end of 2020, just after, it is also noticeable the effect of the vaccination rolled out in France in January 2021 precisely with the oldest people.

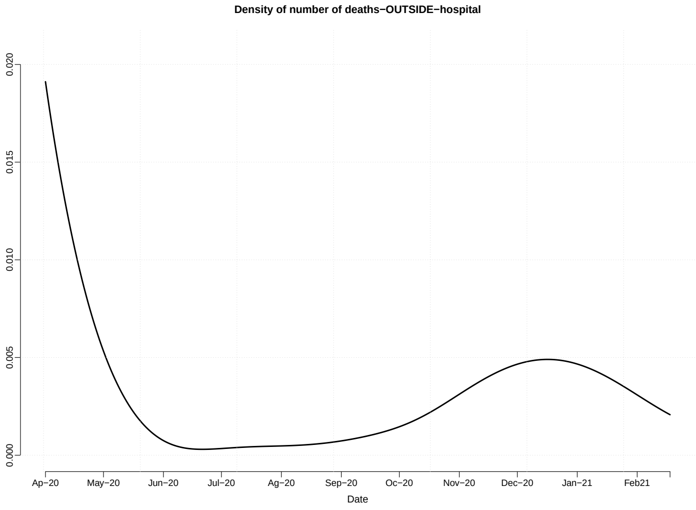

Figure 13: Smooth density estimation of the number of deaths-outside-hospital from April-2020 to February-2021.

6.4.3 Examples of smooth developments of ratio of number of deaths inside versus outside hospital, the Covid-19 case of France

When quantifying the number of deaths outside hospital to be able to find the total number of deaths from the pandemic, then we simply follow the dynamic development of the ratio of people dying inside the hospital versus outside the hospital respectively. We know from similar studies in survival analysis, see Nielsen and Tanggaard (2001), that it is more robust to estimate the numerator and the denominator separately and then divide to get the ratio, then it is to smooth the ratio directly. In Figure 14 we provide the final result of this procedure and we see that while the ratio has stabilized with a little higher probability for dying in hospital compared to outside hospital, then also this ratio has changed significantly during the dynamics of the developing pandemic. In the very beginning double as many people died outside hospital compared to inside hospitals, and this might be down to some of the early problems with keeping care homes free of the infection.

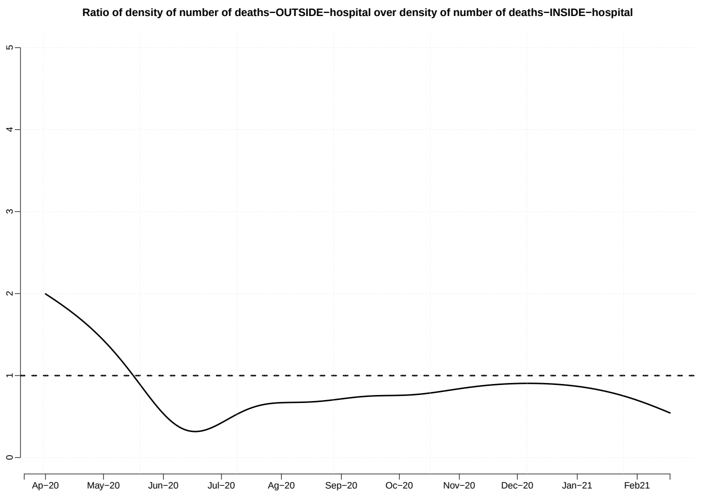

Figure 14: Smooth density estimation of the number of deaths-outside-hospital from April-2020 to February-2021.

6.5 The full system chosen in the Covid-19 case of France

In this section we describe the full system chosen in the Covid-19 case of France (see the full theoretical model in Section 4).The dynamic model will be very dynamic indeed. Even the definition of what we observe might change as the pandemic develops. Our model can be illustrated via the stages represented in Figure 15.

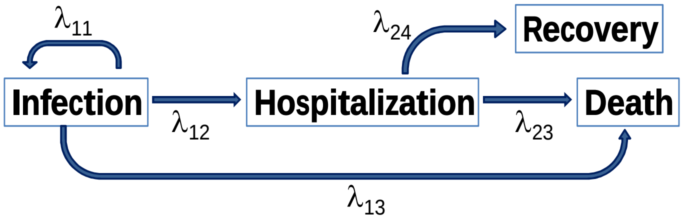

Figure 15: Transition diagram.

In a developing pandemic the first thing that happens is that a few people get infected, further infections are higher when there are many infected, therefore there is a feedback loop from the infected stage to its own stage. The definition of an infected might vary over time. A practical version would be to define infected as those people in the society that have been tested positive for the infection. In the beginning of a pandemic testing is perhaps very limited and one positive test means more than later in the pandemic when testing facilities might be widely available. Also, there might be dynamic in hospitalized. There might be varying decisions over time (also due to seasonal impact) of when a person is hospitalized. Even definition of death might vary over time. While the definition of death itself is robust, it might vary over time what it takes to define one particular death to be due to the pandemic. There is also dynamic in the likelihood of transition from one stage to another. We have used two-dimensional local linear marker dependent hazards, see Nielsen (1998), to capture the transition dynamic between stages. This marker dependent hazard estimator has been manipulated to work for inhomogeneous Poisson processes additional to the independent identically distributed counting process that it was originally designed for. As it has been already mentioned, the two dimensions are duration-in-stage and calendar-time. It is the dependency on calendar-time that provides us with the sought after dynamic.

7 Principles of forecasting. The Covid-19 case of France

The main view of forecasting taken in this paper is related to the considerations of Section 5.3. We want the robust and understandable transitions to play as big a role as possible. When looking at the full system described in Section 3 for the Covid-19 case of France, one can take that point of view that all transitions except that one from the infection indicator to the infection indicator can be described via slowly moving continuous development over time. In other words: except for this one transition, it makes sense - at any given date - to forecast the immediate future structure of these transitions from the immediate past. We are therefore left with the challenge of forecasting the transition from infection indicator to infection indicator. Here we use the C_t explained in Section 5.4.

7.1 Forecasting the infection indicator

It is not an easy task to forecast the infection indicator. Here is a graph of the optimal chosen C_t values for the period may till December 2020.

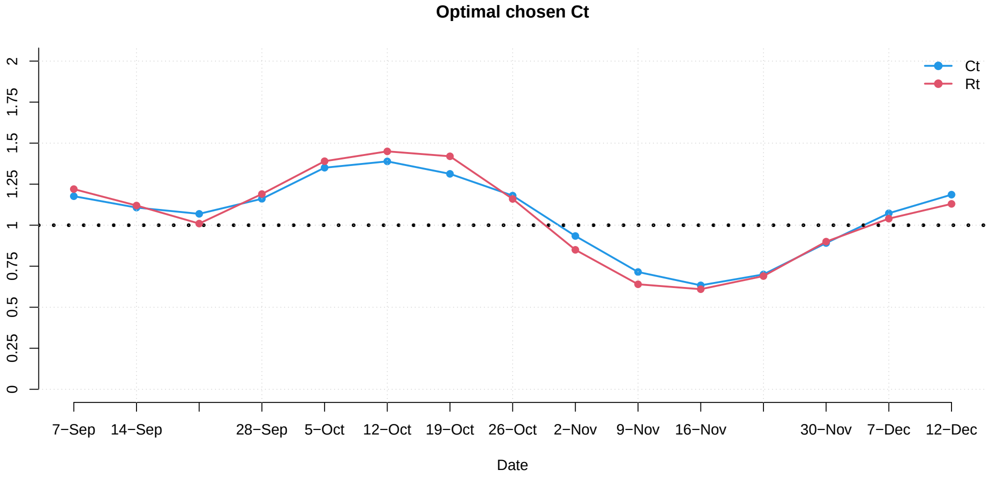

Figure 16: Optimal estimated values of the forecasting constant for one-week predictions

It is clear that $C_t = 1$ is often not a good chose for forecasting the immediate future (see Figures 17 and 18). In other words: the immediate future cannot be forecasted from the immediate past. The point of view taken in this paper is that more information including expert opinion is necessary to make a good chose of C_t at any given date. Notice that C_t is closely related to the much published reproduction number R_t (citation?) on how many new infected one infected individual cause in any given time period. It is reasonable accurate to assume that if the reproduction number R_t is constant then this corresponds to C_t equal to one. If R_t is expected to be bigger in the immediate future then one would need a C_t bigger than one, and vice versa if R_t is expected to be smaller in the immediate future then one would need a C_t smaller than one. We do not have data on the actual R_t and the expected future R_t for France for the period considered. But we think that a reasonable good model for the C_t to use for forecasting could result from a simple regression of best possible C_t down on the actual and the predicted R_t. This is clearly an approach we would recommend during a developing pandemic.

The value of C_t that is plotted on the 7-Sept is the optimal value for predicting new infections in the week 1-Sept to 7-Sept. It is calculated based on the information until 7-Sept. The R_t that is notified on 14-Sept is calculated with information until 7-Sept. Although it is published one week later. We are plotting the R_t series not against the date the have been published, but against the date one week earlier. According to the official website, this number reflects the epidemiological situation one week before notification. That is the number reported on 14-Sept is calculated based on to the information on 7-Sept. In conclusion, every C_t is close to the reported value of R but 7 days later, that is R_{t+7}. There is high correlation between the two indicators.

7.2 Forecasting the 31th of October 2021. The Covid-19 case of France

Figure 17 gives a graphical representation of daily new infected detected since 15-May until 31-October. The black dots give the reported daily numbers until 30-September. The aim in this case is to forecast the number of new infections during the month of October from the information until 30-September. Taking the value $C_t = 1$ we are assuming that the immediate future will behave as the immediate past, then we assume that the rate of infection at the end of the prediction period (31-October) is exactly the same as it was at the end of the observation period (30-Sept). Then we obtain the dash blue line as a forecasting of the number of infected during the month of October. Taking the true number of infected reported in October (grey dots), we can obtain the (infeasible in practice) optimal value of C_t to estimate the infection rate at the end of the prediction period. And then we have estimated that the rate of infection on 31-October is 1.542 times the value of the rate on 30-September. Using linear interpolation we get the infection rate for the whole prediction period which allows to obtain the daily new positives in October (blue dashed line) that best fit the true values. As can be seen, a $C_t = 1$, that is, when no changes are assumed in the behaviour of the infection rate in October with respect to what we had at the end of September, the number of infected will increase as shown by the trend of the red dotted line. However, as we find out when we get the true numbers, the speed of growth of this curve is not sufficient to reproduce the real situation.

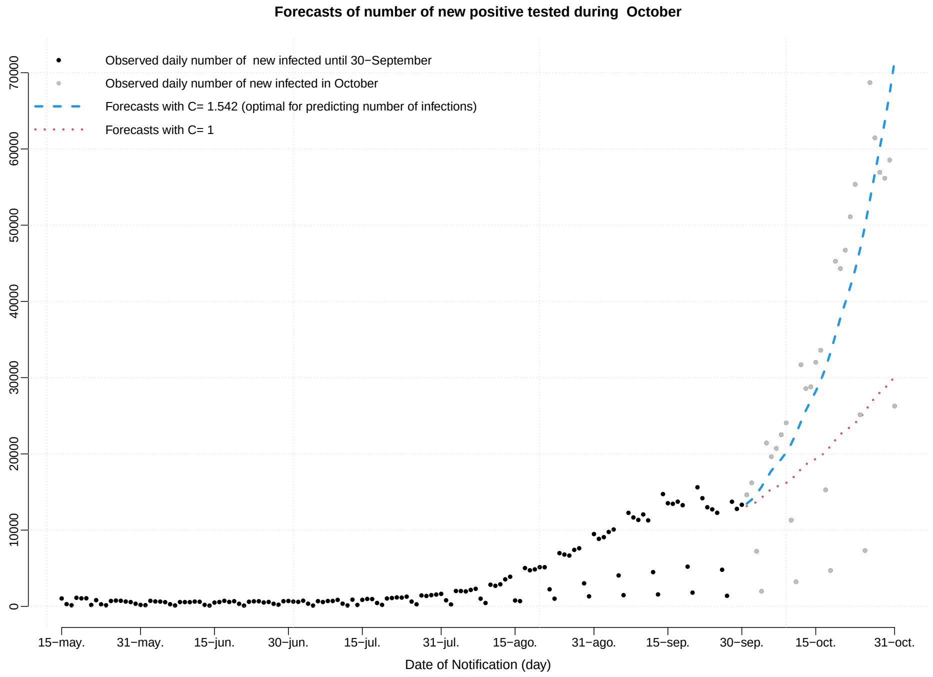

Figure 17: Number of new positives predicted in October using $C_t = 1$ (red dotted line) and using the optimal chosen value for October $C_t = 1.542$ (blue dashed line).

Figure 18 presents a similar study as in Figure 17 but considering this time the problem of predicting the number of new hospitalizations in October on the basis of historical data until September. To obtain the optimal C_t value we do similarly to the previous case. We consider that the rate of infection at the end of the forecasting period is C_t times the value of the infection rate we had estimated for the 30-September. In this case we take the optimal value of C_t by minimizing the error of prediction with respect to the number of hospitalized during the month of October, instead of the number of infected. Then the optimal chosen C_t leads us to a rate of infection on 31-October is 1.783 times the rate of infection on 30-September (blue dashed line). As can be seen, taking $C_t = 1$ in this case leads us to a situation with respect to number of new hospitalized which is very farther from the true situation (red dotted line). To obtain the number of new hospitalized in October we need an additional step, that is, also need to extrapolate the rate of hospitalization estimated with data until 30-September to the month of October. We assume that the hospitalization rate at the end of the forecasting period is exactly the hospitalization rate at the end of the estimation period. Finally, we include in this graph the results obtained considering $C_t = 1.542$ (green dot-dash line), which is the value of the C-constant that minimizes the error with respect to the number of infections in the forecasting period. We pay special attention to this criterion because of its relation with the popular R-number, as it has been discussed in Section 7.1.

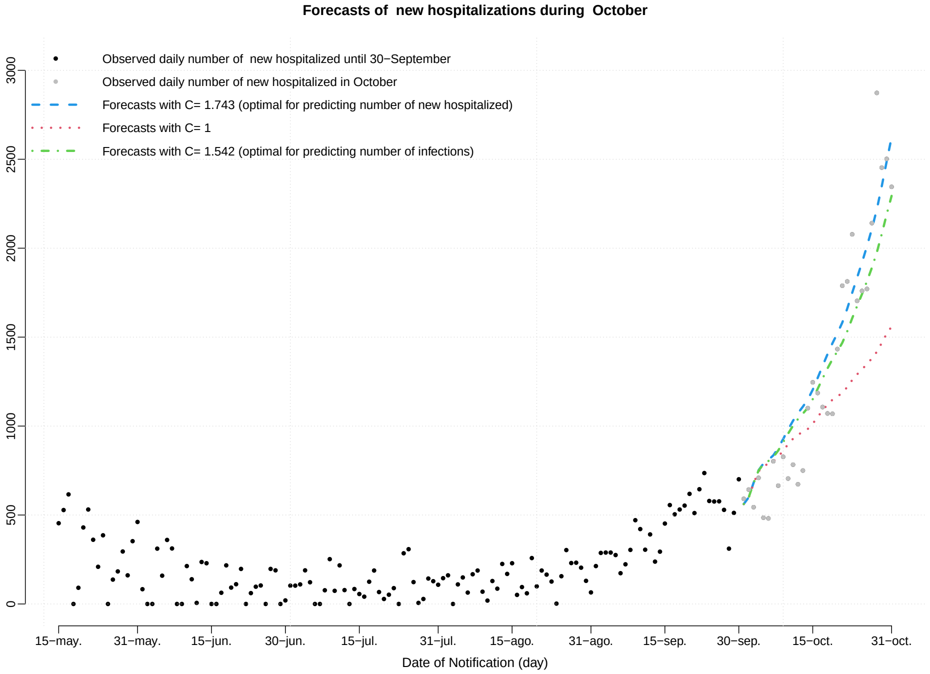

Figure 18: Number of new hospitalized predicted in October using $C_t = 1$ (red dotted line) and using the optimal chosen value with respect to the number of hospitalized in October $C_t = 1.783$ (blue dashed line), and using the optimal chosen value with respect to the number of positives reported in October $C_t = 1.542$ (green dot-dash line).

8 Model formulation

Suppose we are observing n individuals in a time interval $(0, \tau)$.

8.1 Modelling the feedback loop from infected to infected

When modelling the feedback loop from infected to infected we assume that every individual infected give rise to an inhomogeneous Poisson process defined via a two-dimensional forward reaching marker dependent intensity function.

For the ith individual ($i = 1, \dots, n$) and for each $t \in (0, \tau)$, let $Z_{1,i}(t)$ denote the time-dependent covariate taking value 1 when the ith individual has been tested positive for an infectious disease[3] at any time in the interval $(0, t]$; and, 0, otherwise. $Z_{1,i}$ is a one-dimensional marker process with support on the interval $(0, \tau)$.

In general the time-dependent covariate or marker process will be a d-dimensional predictable process. In our case we have $d = 1$, because $Z_{1,i}(t)$ is a process indicating only the date when patient i has tested positive. The marker dependent process $Z_{1,i}$ can be generalized to higher dimensions ($d > 1$) in case other relevant information on patients is incorporated to the model. The marker process Z_{1i} is a predictable, $CADLAG$ covariate and let $F_{1,s}(z) = Pr(Z_{1,i}(s) \leq z)$ denote its conditional distribution function at time s.

Each positive tested patient is likely to be responsible for many other infections in the future. Let $N_{11,i}(t)$ denote the total number of new infections generated by individual i, then, $\{N_{11,1}, \dots, N_{11,n}\}$ are n independent inhomogeneous Poisson processes and $\{(N_{11,1}, Z_{1,1}), \dots, (N_{11,n}, Z_{1,n})\}$ are independent and identically distributed processes. For each individual i, the intensity of the process $N_{11,i}$, $\lambda_{11,i}(t)$, is modelled as depending on the marker process $Z_{1,i}(t)$, by

$$\lambda_{11,i}(t) = \alpha_{11}(Z_{1,i}(t), t), \quad t \in (0, \tau),$$

where $\alpha_{11}(\cdot, \cdot)$ is a deterministic hazard function with no restriction on its functional form.

8.2 Modelling the transition from infected to hospital

When modelling the transition from being infected to entering the hospital, we use the same approach as in Section 8.1 above and use an inhomogeneous Poisson process based on a two-dimensional marker dependent intensity.

[3] The practical application of this paper is focused on COVID-19, but we provide a general methodology.

For the ith individual ($i = 1, \ldots, n$) and for each $t \in (0, \tau)$, let $Z_{1,i}(t)$ denote the marker process associated to the positive test date as in the previous section.
Let $N_{12,i}(t)$ counting the number of hospitalizations of individual i in the interval $(0, t)$. We assume that $\{N_{12,1}, \ldots, N_{12,n}\}$ are n independent inhomogeneous Poisson processes and that $\{(N_{12,1}, Z_{1,1}), \ldots, (N_{12,n}, Z_{1,n})\}$ are independent and identically distributed processes.

For each individual i, the intensity of the process $N_{12,i}$, $\lambda_{12,i}(t)$, is modelled as depending on the marker process $Z_{1,i}(t)$, by

$$\lambda_{12,i}(t) = \alpha_{12}(Z_{1,i}(t), t), \quad t \in (0, \tau),$$

where $\alpha_{12}(\cdot, \cdot)$ is a deterministic hazard function with no restriction on its functional form.

8.3 Modelling transition from hospitalized to death or hospitalized to recovered

When modelling transition from hospitalized to death or hospitalized to recovered, we use the standard counting process set-up as defined in Nielsen (1998). Then we work with duration modelling from entering hospital to leaving hospital either as recovered or a dead by infection.
For the ith individual ($i = 1, \ldots, n$) and for each $t \in (0, \tau)$, let $Z_{2,i}(t)$ denote the variable taking value 1 when the ith individual has been hospitalized on a date prior or equal to t; and, 0, otherwise. Let $N_{2,i}(t)$ take value 1 if the ith individual leaves the hospital (due to death or recovery, whichever comes first) in the interval $(0, t]$; and, 0 otherwise. We assume that $N_{2,i}$ is a one-dimensional counting process with respect to an increasing, right continuous, complete filtration $\mathcal{F}_{2,t}$, $t \in (0, \tau)$, i.e. it obeys *les conditions habituelles*; and again we assume Aalen's multiplicative model for the intensity function, that is,

$$\lambda_{2,i}(t) = \alpha_2(Z_{2,i}(t), t) Y_{2,i}(t), \quad t \in (0, \tau),$$

where $\alpha_2(\cdot, \cdot)$ is a deterministic hazard with no restriction on its functional form. $Y_{2,i}$ is a predictable process taking values in $\{0, 1\}$, indicating (by the value 1) when the ith individual is in the hospital and at risk. We assume that $E[Y_{2,i}(s)] = \gamma_2(s)$, where $\gamma_2(\cdot)$ is continuous and the marker $Z_{2,i}(s)$ is only observed for those s where $Y_{2,i}(s) = 1$.

We assume that $\{(N_{2,1}, Z_{2,1}, Y_{2,1}), \ldots, (N_{2,n}, Z_{2,n}, Y_{2,n})\}$ are independent and identically distributed and $\mathcal{F}_{2,t} = \sigma(\mathbf{N}_2(s), \mathbf{Z}_2(s), \mathbf{Y}_2(s); s \leq t)$, where $\mathbf{N}_2 = (N_{2,1}, \ldots, N_{2,n})$; $\mathbf{Y}_2 = (Y_{2,1}, \ldots, Y_{2,n})$ and $\mathbf{Z}_2 = (Z_{2,1} \ldots, Z_{2,n})$.

One individual might leave hospital due to death or recovery, whichever occurs first. Then we can define the corresponding intensity of leaving the hospital by one specific cause as it is represented in Figure 15. Then it can be written

$$\begin{aligned} \lambda_{23,i}(t) &= \alpha_{23}(Z_{2,i}(t), t)Y_{2,i}(t), \quad \text{and,} \\ \lambda_{24,i}(t) &= \alpha_{24}(Z_{2,i}(t), t)Y_{2,i}(t), \end{aligned}$$

for $t \in (0, \tau)$ and with $\alpha_2 = \alpha_{23} + \alpha_{24}$.

Appendix A. The algorithm

Assume we observe daily counts of occurrences and exposures in a grid of time points $\{1, 2, \ldots, M\}$. Denote by z the day of onset, x the day of outcome (non-recurrent or recurrent event) and $d = x - z + 1$ the number of days until an outcome is registered. Define $O_{x,d}$ as the number of subjects that entered the system at time $z = x - d + 1$ and have been there for a time equal to d, with $1 \leq z < x \leq M$. Denote by $E_{x,d}$ the number of arrivals registered at time $z = x - d + 1$ and have been there for exactly d days. Assume that there are no observations before day 1. The sequence $\{O_{x,d}, E_{x,d}; 1 \leq z \leq M, 1 \leq d \leq M - z + 1\}$ is observable when full information is available. We consider the case of partial information when only the marginal counts $O_x = \sum_{d=1}^{x} O_{x,d}$ for occurrences, and $E_x = \sum_{d=1}^{x} E_{x,d}$ are available, where x denotes the outcome notification day.

The algorithm is closely related to that of Gamiz *et al.* (2022) and consists of estimating $\alpha(z, d)$ by an iterative procedure.

We consider to different situations.

1. *Survival model*: Subjects abandon the system at the occurrence of the (non-recurrent) final event.

   In this first situation of survival analysis, we consider two different types of departures for the subjects entering the system. In our practical application, an arrival means a patient entering the hospital and the outcome can be the death of the patient or hospital discharge due to recovery, whichever occurs first. Then, the algorithm consists of two main steps.

   Get initial values $\hat{O}_{x,d}^{(0)}$ and $\hat{E}_{x,d}^{(0)}$ from an initial choice for $\widehat{\alpha}^{(0)}(z,d)$, e.g. $\widehat{\alpha}^{(0)}(z,d) \equiv \widehat{\alpha}^{(0)}(d)$ from an Exponential distribution, with $1 \leq z, d \leq M$, $x = z + d - 1$.

   The r-th iteration cycle of the first step of the algorithm consists of the following steps:

   (i) Put $d = 1$ ($x = z + d - 1$). Define $\hat{O}_{x,d}^{(r)}$ as the mean value of a Binomial distribution with parameters $\hat{E}_{x,d}^{(r)}$ and probability $\widehat{\alpha}^{(r)}(z,d)$.

   (ii) Define $\hat{E}_{x+1,d+1}^{(r)} = \hat{E}_{x,d}^{(r)} - \hat{O}_{x,d}^{(r)}$, and then define $\hat{O}_{x+1,d+1}^{(r)}$ as the mean value of a Binomial with parameters $\hat{E}_{x+1,d+1}^{(r)}$ and $\widehat{\alpha}^{(r)}(z,d+1)$, repeat for $d = 3, \ldots$.

   The following sequences of occurrences and exposure are obtained:

$$\left\{ \left(\hat{O}_{x,1}^{(r)}, \hat{E}_{x,1}^{(r)} \right), \ldots, \left(\hat{O}_{M,M-x+1}^{(r)}, \hat{E}_{M,M-x+1}^{(r)} \right) \right\}$$

   (iii) Define

$$q_{x,d}^{(r)} = \frac{\hat{O}_{x,d}^{(r)}}{\sum_{d'=1}^{M} \hat{O}_{x,d'}^{(r)}} \quad \text{and} \quad h_{x,d}^{(r)} = \frac{\hat{E}_{x,d}^{(r)}}{\sum_{d'=1}^{M} \hat{E}_{x,d'}^{(r)}}$$

   (iv) Put $r = r + 1$ and update the occurrences and exposures:

$$\hat{O}_{x,d}^{(r)} = q_{x,d}^{(r-1)} O_x, \text{ and } \hat{E}_{x,d}^{(r)} = h_{x,d}^{(r-1)} E_x.$$

   (v) Re-arrange the estimated exposure and occurrences in a matrix where rows correspond to entry day (z) and columns to duration (d), that is, for $1 \leq z \leq M$ and $1 \leq d \leq M - z + 1$ define $\hat{O}_{z,d}^{*} = \hat{O}_{z+d-1,d}$, and $\hat{E}_{z,d}^{*} = \hat{E}_{z+d-1,d}$, and use these matrices to estimate the two-dimensional local-linear hazard (Nielsen (1998), Gamiz *et al.*(2013):

$$\widehat{\alpha}_{\mathrm{b}}(z_0, d_0) = \frac{\sum_{z=1}^{M} \sum_{d=1}^{M-z+1} \bar{\mathcal{K}}_{\mathrm{b}}(z_0 - z, d_0 - d)\hat{O}^{*}_{z,d}}{\sum_{z=1}^{M} \sum_{d=1}^{M-z+1} \bar{\mathcal{K}}_{\mathrm{b}}(z_0 - z, d_0 - d)\hat{E}^{*}_{z,d}}$$

where b is a two-dimensional bandwidth, and $\bar{\mathcal{K}}$, is the generalization of the one-dimensional local-linear kernel as described in Section 3.

(vi) Repeat until convergence.

In the second step, the final hazard estimator for duration, $\widehat{\alpha}$, is split into hazards for duration due to one of two possible types of outcome, for example hazards due to death ($\widehat{\alpha}^D$) and hazards due to recovery ($\widehat{\alpha}^R$), as follows: For deaths we have

$$\widehat{\alpha}^{D}_{\mathrm{b}}(z_0, d_0) = \frac{\sum_{z=1}^{M} \sum_{d=1}^{M-z+1} \bar{\mathcal{K}}_{\mathrm{b}}(z_0 - z, d_0 - d)\hat{O}^{*D}_{z,d}}{\sum_{z=1}^{M} \sum_{d=1}^{M-z+1} \bar{\mathcal{K}}_{\mathrm{b}}(z_0 - z, d_0 - d)\hat{E}^{*}_{z,d}},$$

where $\hat{O}^{*D}_{z,d} = \hat{O}^{D}_{z+d-1,d}$, with $\hat{O}^{D}_{x,d} = \hat{q}_{x,d}\, O^{D}_{x}$, and O^{D}_{x} being the total number of deaths registered on the day x. And for recoveries we have

$$\widehat{\alpha}^{R}_{\mathrm{b}}(z_0, d_0) = \frac{\sum_{z=1}^{M} \sum_{d=1}^{M-z+1} \bar{\mathcal{K}}_{\mathrm{b}}(z_0 - z, d_0 - d)\hat{O}^{*R}_{z,d}}{\sum_{z=1}^{M} \sum_{d=1}^{M-z+1} \bar{\mathcal{K}}_{\mathrm{b}}(z_0 - z, d_0 - d)\hat{E}^{*}_{z,d}},$$

where $\hat{O}^{*D}_{z,d} = \hat{O}^{D}_{z+d-1,d}$, with $\hat{O}^{D}_{x,d} = \hat{q}_{x,d}\, O^{D}_{x}$, and O^{R}_{x} is the total number of recoveries observed on the day x, $1 \leq x \leq M$.

2. *The inhomogeneous Poisson model: recurrent events.*

Appendix B. Simulations