WPOM

Working Papers on Operations Management.
Vol. 12, Nº1 (76-115)
ISSN: 1989-9068

http://dx.doi.org/10.4995/wpom.15332

A data generator for covid-19 patients' care requirements inside hospitals

Juan A. Marin-Garcia[a], Angel Ruiz[b], Julien Maheut [c] and Jose P. Garcia-Sabater[d]

[a] ROGLE. Dpto. de Organización de Empresas. Universitat Politècnica de València (Spain) jamarin@omp.upv.es , [b]Département d'opérations et systèmes de décision , FSA ULaval(Canada). Angel.Ruiz@osd.ulaval.ca, [c] ROGLE. Dpto. de Organización de Empresas. Universitat Politècnica de València (Spain) juma2@upvnet.upv.es and [d] ROGLE. Dpto. de Organización de Empresas. Universitat Politècnica de València (Spain) jpgarcia@omp.upv.es.

Recibido: 2021-03-27 Aceptado: 2021-05-24

To cite this article: Marin-Garcia, J.A.; Ruiz, A.; Maheu, J.; Garcia-Sabater, J.P. (2021). *A data generator for covid-19 patients' care requirements inside hospitals. WPOM-Working Papers on Operations Management,* 12 (1), 76-115. doi: https://doi.org/10.4995/wpom.15332

Abstract

A Spanish version of the article is provided (see section before references). This paper presents the generation of a plausible data set related to the needs of COVID-19 patients with severe or critical symptoms. Possible illness' stages were proposed within the context of medical knowledge as of January 2021. The parameters chosen in this data set were customized to fit the population data of the Valencia region (Spain) with approximately 2.5 million inhabitants. They were based on the evolution of the pandemic between September 2020 and March 2021, a period that included two complete waves of the pandemic.

Contrary to expectation and despite the European and national transparency laws (BOE-A-2013-12887, 2013; European Parliament and Council of the European Union, 2019), the actual COVID-19 pandemic-related data, at least in Spain, took considerable time to be updated and made available (usually a week or more). Moreover, some relevant data necessary to develop and validate hospital bed management models were not publicly accessible. This was either because these data were not collected, because public agencies failed to make them public (despite having them indexed in their databases), the data were processed within indicators and not shown as raw data, or they simply published the data in a format that was difficult to process (e.g., PDF image documents versus CSV tables). Despite the potential of hospital information systems, there were still data that were not adequately captured within these systems.

Moreover, the data collected in a hospital depends on the strategies and practices specific to that hospital or health system. This limits the generalization of "real" data, and it encourages working with "realistic" or plausible data that are clean of interactions with local variables or decisions (Gunal, 2012; Marin-Garcia et al., 2020). Besides, one can

parameterize the model and define the data structure that would be necessary to run the model without delaying till the real data become available. Conversely, plausible data sets can be generated from publicly available information and, later, when real data become available, the accuracy of the model can be evaluated (Garcia-Sabater and Maheut, 2021).

This work opens lines of future research, both theoretical and practical. From a theoretical point of view, it would be interesting to develop machine learning tools that, by analyzing specific data samples in real hospitals, can identify the parameters necessary for the automatic prototyping of generators adapted to each hospital. Regarding the lines of research applied, it is evident that the formalism proposed for the generation of sound patients is not limited to patients affected by SARS-CoV-2 infection. The generation of heterogeneous patients can represent the needs of a specific population and serve as a basis for studying complex health service delivery systems.

Keywords: *data paper; simulated data set; covid-19; hospital; bed management; healthcare; operations management*

Introduction

From the beginning of the health crisis associated with COVID-19, each healthcare system had to deal with fluctuating healthcare requirements, facing pandemic waves of varying amplitude and duration, conditioned in part by the application of healthcare restrictions and protocols. From the beginning of the crisis, many scientists worked on designing solutions to mitigate the effects of COVID-19. These ranged from the development of vaccines to tools for forecasting infections and the impact of political mitigation measures using advanced artificial intelligence techniques. However, the management of hospitals and their resources did not seem to receive as much attention (Epstein & Dexter, 2020). Each country, each region, and even each hospital came to manage its critical resources (beds and healthcare personnel) locally, without the existence of coordination mechanisms and tools to anticipate and mitigate the consequences of the waves, which were usually produced by the aggregation of local infection clusters.

Hospital bed management is a concrete application of a generic problem of capacity management (Claudio et al., 2021; Garcia-Sabater et al., 2020; Lagarda-Leyva & Ruiz, 2019; Marin-Garcia et al., 2019; Nino et al., 2021; Xia & Sun, 2013). For this purpose, it is possible to use operations management tools, and more specifically design, planning, and control or process-improvement tools. In this sense, discrete event-based simulation is a tool to support hospital management decision making (Gunal, 2012; Marin-Garcia, Garcia-Sabater, et al., 2020). With a process simulator, it is possible to facilitate adequate planning of healthcare resources and to anticipate, or at least to mitigate, situations in which some health centers cannot attend patients due to saturation or system collapse while other centers have idle resources (Romeo Casabona & Urruela Mora, 2020). It can also be used to empirically determine occupancy thresholds before diverting patients between hospitals and thus avoid transferring patients in more advanced disease stages and, therefore, requiring a much more complicated, risky, and costly transfer.

Note this approach to the bed management problem does not attempt to predict whether a particular person will recover, progress to a more serious condition, or die, nor does it predict the exact number of days a particular patient will be in each stage. The objective is to model the care required and the associated processes within the hospital for COVID-19 patients in order to predict with sufficient reliability the overall daily occupancy of hospital beds, the use of Non-Invasive Mechanical Ventilation (NIMV) equipment, the occupancy of beds in Intensive Care Units (ICU), the need for medical staff, and the need for patient referral within the chosen study area. In this sense, the prediction consists of estimating probabilities for different occupancy rates on certain days in the future or whether, given a current occupancy and taking into account the rate of COVID-19 admissions into the hospital, when there will no longer be free beds available in hospitalization or ICU requiring patient referrals or enabling more capacity in the COVID-19 patient circuit.

One of the problems to be solved is the modeling of statistical distributions, both of the time required for care and of the different stages COVID-19 patients may go through, which requires a significant volume of data (Gunal, 2012). Contrary to expectations and in spite of the European and national transparency laws (BOE-A-2013-12887, 2013; Parlamento Europeo y del Consejo de la Unión Europea, 2019), actual data related to the COVID-19 pandemic, at least in Spain, took a long time to be updated and made available (usually a week or ten days). In addition, some data relevant for working with hospital bed management models were not publicly accessible. This occurred for various reasons including that these data were not collected, public agencies did not offer the data (despite having them indexed in their internal databases), public agencies offered the data processed into indicators and did not show the raw data, or they simply published the data in a format that was difficult to reuse (e.g., PDF image documents versus CSV tables).

On the other hand, despite the fact that hospital information systems were quite powerful, there were still data that were not adequately collected within these systems. These were data that were not recorded in the process of admission or during the treatment of patients or, if they were recorded, they were recorded in fields or in a non-standardized format. This forced manual extraction and filtering of the information, which potentially generated errors and prevented easy access to the data.

Another problem, although not the least, lay in the existing interdependence between the data collected in a hospital and the strategies and practices of that hospital or health system. Thus, when we looked at the data, what we saw in effect was the result of a series of policies applied. Some of these policies were explicit while others were implicit and, in many cases, unobservable. This effect limited the generalizability of the "real" data and, in many cases, it was much more interesting to work with "realistic" or plausible data that were free of interactions with local variables or decisions.

For these reasons and for the simulation of internal processes of the treatment of COVID-19 patients in hospitals, it may be appropriate to do so with realistic rather than real data (Gunal, 2012). It would be shortsighted to delay the creation of a model while waiting for the actual data (Garcia-Sabater & Maheut, 2021). One can parameterize the model and define the data structure that would be necessary to run the model with real data. Conversely, plausible data sets can be generated from publicly available information and, later, when real data become available, the accuracy of the model can be evaluated (Garcia-Sabater & Maheut, 2021).

In order to address this situation, we proposed an algorithm for the generation of plausible data related to the health care needs of symptomatic COVID-19 patients (those displaying severe or critical symptoms).

We used this algorithm to generate a data set to serve as a test bed for simulation models to be generated in the future. In this way, we could test the effect of different decisions related to hospital bed management (e.g., triage or patient discharge levels, referrals, capacity increase, etc.) on the trajectories and outcomes of the patients generated. In addition, this data set facilitated checking whether, for the generation of discrete event simulation models for hospital bed management, the data tables generated were sufficient or whether additional variables not yet considered in the proposed data sets were necessary.

The data set generated in this paper facilitates rapid generation of new research, reproducibility of research, and validation of results (Marin-Garcia, 2015; Roa-Martínez et al., 2017). Thanks to the reuse of data or the creation of new data sets by means of the script we provide it is possible to compare the goodness of fit of different models or to draw new conclusions by re-analyzing the same data set using alternative techniques or approaches.

Method

Data generation was conducted using RStudio (RStudio Team, 2020) and various R packages (Comtois, 2021; R Core Team, 2020; Revelle, 2021; Ruckdeschel et al., 2006; Schauberger & Walker, 2020; Venables & Ripley, 2002; Wickham, 2007, 2011; Wu et al., 2020): MASS, summarytools, stats, psych, plyr, dplyr, distr, ExtDist, openxlsx, and reshape2.

Ethical Statement

No private personal data were handled nor was there a need for human participants. Consequently, ethical review and approval were not required for the study in accordance with local legislation and institutional requirements.

Objective

The fundamental objective of the proposed method was to create a data set that represents the clinical needs of COVID-19 patients admitted to hospitals in a geographic area over a time period consisting of two complete pandemic waves. It considers stages possible based on medical knowledge available as of January 2021. If new treatment stages are incorporated in the future or if it is considered appropriate to eliminate any of those included in the proposed model, the method can be modified to adjust it to the new reality.

The parameters chosen in this data paper were customized to fit the population values of the Valencia region (Spain), which had approximately 2.5 million inhabitants, representing the evolution of the pandemic between September 2020 and March 2021. In any case, the time horizon covered, the intensity of the number of hospital admissions in the period, and the incidence curve (sharper or flatter) are fully customizable to fit any other scenario.

Preliminary Assumptions for Data Generation

This work focused on the requirement paths of patients. Specifically, the sequence of symptom evolution that caused a patient to require certain care needs. Therefore, the trajectories (scheduling, coordination, interaction, and resource allocation of all the necessary care steps within a health center) were not addressed (Alexander, 2007; Corbin & Strauss, 1988; Pinaire et al., 2017; Unroe et al., 2010), as the trajectories were conditioned not only by the needs of the patients but also by the availability of resources.

The data generation algorithm modeled the clinical resource needs of the patient, not the resources actually allocated to the patient after applying a particular triage policy in a hospital. To generate the patient data, we relied on information provided by internists, intensivists, medical area directors, and publicly available information as of the end of December 2020 as summarized below:

1. There was no treatment that could change the course of the stages that a patient went through (Plaza, 2021). That is, there was only one specific treatment for COVID-19, providing oxygen, which was supplied to patients who needed it. Oxygen can be supplied in three degrees of intensity including oxygenation (mask in a normal hospital bed), non-invasive forced ventilation (NIMV), and invasive forced ventilation (IMV) (Daniel et al., 2021; European center for disease prevention and control, 2020; Fowler et al., 2020; Manninen, 2020; Marin-Garcia, Garcia-Sabater, et al., 2020; Winck & Scala, 2021)
2. Exposure to the virus did not always guarantee infection. However, once infected (exposed to a sufficient dose to develop the disease), each person developed a trajectory that was predetermined (although unknown) at the moment of infection. The trajectory of the patients was a set of stages, in a predefined order, with variable duration times in each of them (the time may be zero for some) (Fowler et al., 2020; Wong et al., 2020).
3. The disease had a progressive and variable speed process (although in most cases it was slow, the evolution from one state to another usually taking several days). No patient needed to be admitted to the ICU without first needing to be admitted to a hospital for observation (Belciug et al., 2020; Olivieri et al., 2021; Stang et al., 2020), and should also not have been discharged post-ICU without undergoing a period of observation in inpatient beds (Castelnuovo et al., 2020; ECDP, 2020).
4. Each row in the model's output file represented one patient. For the distribution of gender, age, and comorbidity, the values of the other cells of the patient row were not taken into account (in future versions the model can be improved by representing the relationship between the variables).

If any of these assumptions are found to be incorrect in any given context, parameters or constraints can be added in the future to more adequately represent a particular situation.

Patient Care Requirements

Figure 1 shows schematically the evolution of patient symptoms and the type of care required at each point (Daniel et al., 2021; Fowler et al., 2020; Winck & Scala, 2021). Of those infected with SARS-CoV-2, a small percentage will develop severe or critical COVID-19 symptoms. Ideally, these individuals will require specific medical observation, usually with admission to a hospital for care and follow-up. The standard provision of all hospital beds in the Spanish public health system allowed for oxygen therapy without forced ventilation if necessary. A portion of the patients will improve after a few days and will be able to complete their recovery at home when their symptoms are moderate or mild (R1). However,

others will worsen and require NIMV (Brochard, 2003; Popat & Jones, 2012; Ruza, 2008; Wiersema, 2007) which could be applied in normal rooms that have undergone a small low-cost adaptation. These adaptations would relieve the pressure on ICU beds as the patients do not yet require the special conditions of an ICU. After a period of treatment with forced ventilation, a portion of the patients will improve and will be able to return to a normal hospital bed where they will continue their recovery before being discharged to home care (R2). The remainder will have more severe symptoms and will require IMV (Brochard, 2003; Buckley & Gillham, 2007; Popat & Jones, 2012; Ruza, 2008), which can only be administered in the ICU or assimilated beds (e.g., pediatric ICU, operating rooms, emergency ICUs). Patients who recover after the ICU stay will move to a hospital bed (post-ICU) to complete recovery before being discharged home (R3). The remainder will die (R4).

Figure 1. Patient requirements process flow

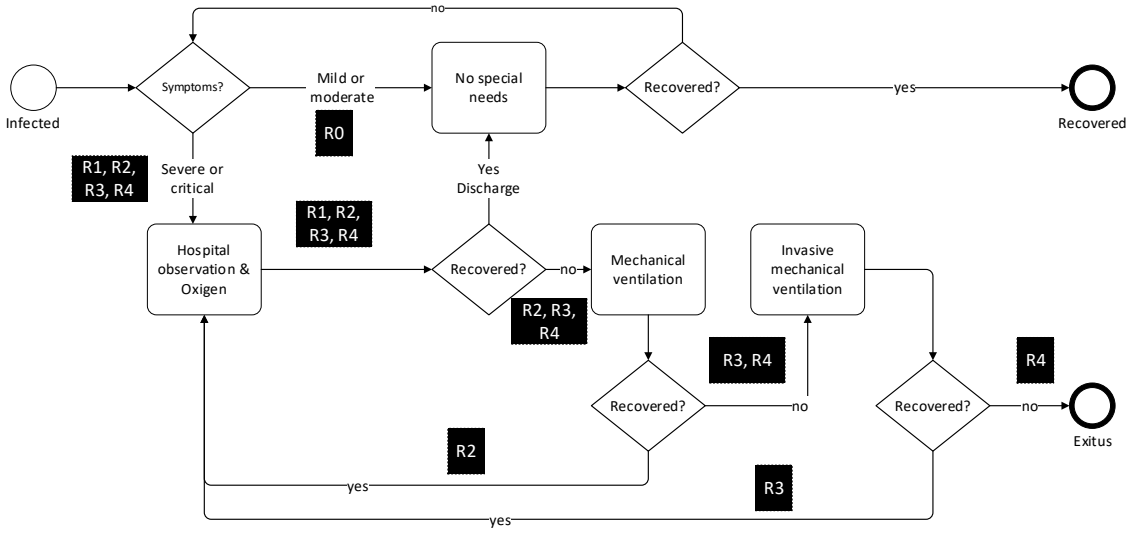


Table 1 shows the breakdown of the flow diagram in Figure 1 into the different care requirement paths considered in this work. The generator focuses on patients requiring hospitalization classed as severe and critical (Fowler et al., 2020), i.e., paths R1, R2, R3, and R4. In each requirement path there may be different "patient families" if the duration of resource use is different according to some set of identifiable variables. If not, there will simply be a dispersion in the requirement path data that cannot be explained/grouped by families. In this paper we considered patient families grouped by age ranges.

Table 1. Patien care requirements paths

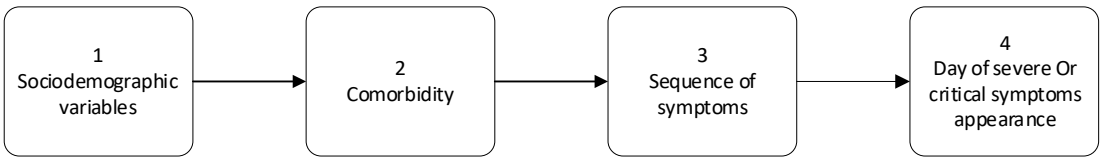
Estates	Requirement path	Observation & non forced ventilation	Mechanical ventilation	Invasive mechanical ventilation	Observation	Recovered
Mild or moderate	R0	0	0	0	0	1
Severe-critical	R1	1	0	0	0	1
	R2	1	1	0	1	1
	R3	1	1	1	1	1
	R4	1	1	1	0	0

Procedure

We used official figures from the Spanish National Center of Epidemiology (CNE) (CNE -Centro Nacional de Epidemiología, 2020) for total cases, hospitalized patients, ICU patients, and deaths by day, province, age group, and gender between January 1, 2020 and March 24, 2021. This process was imperfect, e.g., the data set with cases took time to be updated, there were possible minor inconsistencies between regional protocols, and the CNE data only gave totals for groups over time forcing the use of different sources to make assumptions about the time between steps in requirement paths. Finally, the CNE data did not include complete information about comorbidity, so we made assumptions based on other sources.

The process of data generation is carried out in four steps (Figure 2).

Figure 2. Four steps for data generation



Step 1

In Step 1, the sociodemographic variables are created. Our simulation was only based on severe or critically ill patients requiring hospitalization. The gender and age distributions of these patients may differ from those of the general population in Valencia province (Generalitat Valenciana. Conselleria de Sanitat Universal i Salut Pública, 2019; Generalitat Valenciana, 2018). The number of hospitalized men was 1.2 times the number of hospitalized women (men, 54.5%; women, 45.4%) (CNE -Centro Nacional de Epidemiología, 2020). For age, we worked with the age ranges used by the CNE, specifically, 0.4% of cases under 10 years of age, 0.6% between 10 and 19 years, 2.4% between 20 and 29 years, 4.6% between 30 and 39 years, 10.4% between 40 and 49 years, 15.3% between 50 and 59 years, 18.0% between 60 and 69 years, 20.2% between 70 and 79 years, and the remainder 80 years and above. (CNE - Centro Nacional de Epidemiología - https://cnecovid.isciii.es/).

Step 2

In Step 2, comorbidity is randomly distributed among the generated sample to respect an incidence similar to that of the COVID-19 patient sample. (CNE -Centro Nacional de Epidemiología, 2020). Data for Spain were available only up to Report 32 dated May 21, 2020, after which these data were not reported. Nor were we able to locate information on the simultaneous prevalence of comorbidities or the relationships of these with gender and age in the general population (Posso et al., 2020). For this reason, they were randomly distributed to maintain a proportion of cases where the prevalence was 28.5% of heart conditions such as heart failure, coronary artery disease, or cardiomyopathies; 11.6% of chronic obstructive pulmonary disease; 17.7% of type 2 diabetes mellitus; 12.9% of hypertension; 2.6% of chronic kidney disease; and 3.3% of cancer. The CNE did not provide data on other risk factors (e.g., 16.9% of obesity (BMI > 30 kg/m2), 23.0% of smoking), so they were extracted from other sources (Ministerio De Sanidad, Servicios Sociales e Igualdad, 2017). In this study we did not include pregnancy as a risk factor (no comorbidity). In the future

this variable can be included by assigning cases only to women in the fertile age range. In future research, we can also improve the generation of cases by taking into account the incidence of comorbidity according to sociodemographic variables and the simultaneous presence of several comorbidities (Zheng et al., 2020). In addition, the values can be adapted to other populations by incorporating information from other sources such as https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/evidence-table.html.

Step 3

In Step 3, each patient is assigned a trajectory according to the risks derived from their demographic and comorbidity variables. Despite having formally requested them from various organizations, it was not possible to obtain public data allowing us to establish the percentage of patients in each trajectory by age range. Therefore, we estimated these percentages from the only readily available data source (CNE - Centro Nacional de Epidemiología - https://cnecovid.isciii.es/). We used data from across Spain (Figure 4 and Table 2). We did not use data exclusively from Valencia province (Figure 3) because there were few cases of patients under 30 years of age thus rendering the parameter estimates unreliable. The CNE data did not adequately represent the trajectories of people over 60 years of age. It may misleadingly appear that the incidence of Severe Acute Respiratory Syndrome (SARS) decreases after this age. However, the reality was the opposite, this incidence grew exponentially, but the application of medical care protocols meant most elderly patients did not receive ICU treatment even if they were in need of IVM. The reason was the probability of surviving such an aggressive intervention was low (practically zero in the most elderly patients).

Figure 3. Evolution of variables in Valencia province

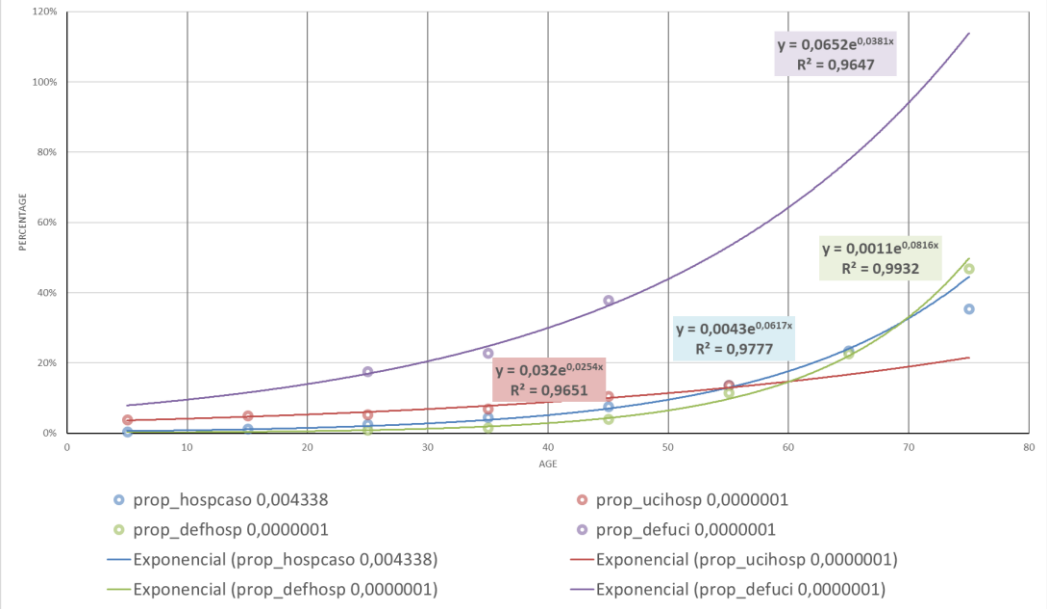


Figure 4 Evolution of variables in Spain

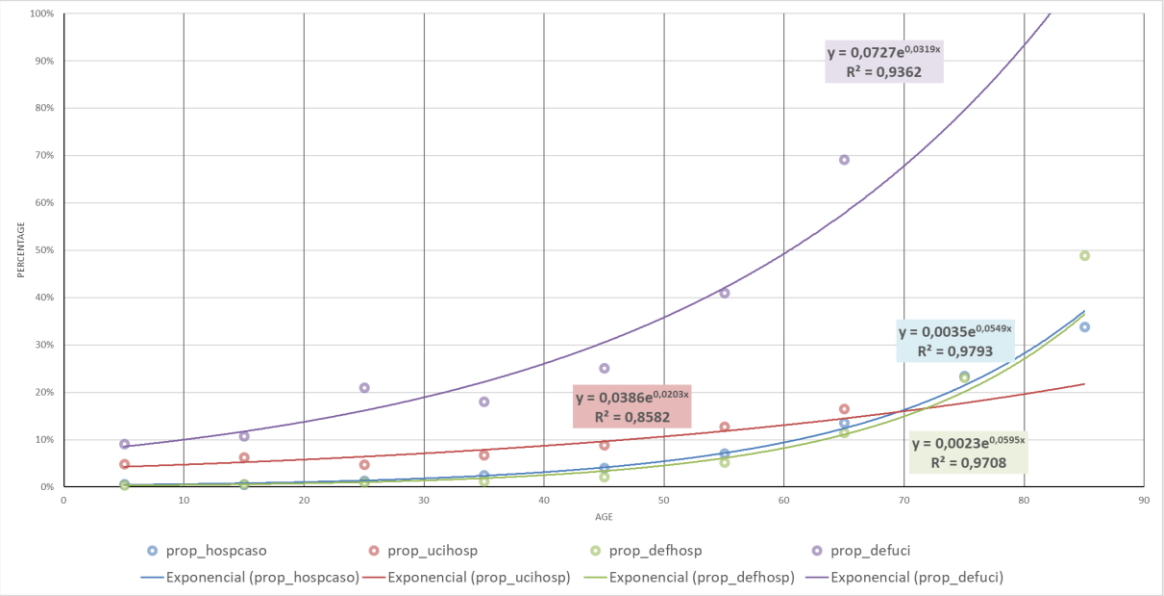


Table 2 Estimated ICU patients based in Spanish figures

B	C	D	E	F	G	H	I	J	K	L
Age Range	Infected	Hopitaliced	ICU	Exitus	Hosp per infected	ICU per Hospitalized	Exit per Hospitalized	Exit per ICU	Estimated ICU per Hospitalized	Estimated Exitus per ICU
5	202429	1351	66	6	0,006674	0,048853	0,004441	0,090909	0,04272363	0,08527159
15	324036	1624	102	11	0,005012	0,062808	0,006773	0,107843	0,05233955	0,11731251
25	357745	4759	228	48	0,013303	0,047909	0,010086	0,210526	0,06411974	0,16139284
35	378003	9498	640	116	0,025127	0,067383	0,012213	0,18125	0,07855133	0,22203641
45	468477	18976	1682	422	0,040506	0,088638	0,022239	0,250892	0,09623107	0,30546689
55	413136	29652	3776	1549	0,071773	0,127344	0,052239	0,410222	0,11789003	0,42024648
65	263693	35876	5944	4108	0,136052	0,165682	0,114506	0,691117	0,14442384	0,57815464
75	168258	39438	5280	9141	0,23439	Not available	0,231782	Not available	0,17692966	0,79539702
85	172818	58510	1038	28635	0,338564	Not available	0,489404	Not available	0,21675166	0,90(1)
(1) Value imputed manually since the estimate shows a value outside the possible range									=0,0386* EXP(0,0203* AgeRange)	=0,0727* EXP(0,0319* AgeRange)

In order to establish the parameters of the percentage of people on the R3 and R4 paths, the curve of points corresponding to the proportion of ICU patients per number hospitalized and the percentage who died (exitus) per number of ICU patients were fitted with an exponential (Figure 4 and Table 2). These calculated values (Columns K and L in Table 2) were subtracted from the flow of hospitalized patients to calculate the weight of each of the paths representing the evolution of symptoms.

Taking into account the previous assumption number three, path R4 will have a prevalence equal to the proportion of deceased persons in relation to the total number of hospitalized persons. These patients will have required mechanical ventilation (whether or not it was administered).

The set of patients on paths R2 and R3 is comprised of those who required mechanical ventilation. This figure could be estimated from the persons who required admission to the ICU (in Spain there were few hospitals that offered NIMV in the first 10 to 12 months of the epidemic). However, Table 2 clearly shows the trend in ICU admissions stagnated and decreased from the age range of 65 years onwards, while the number of patients who died (exitus) grew exponentially. This trend indicated, after a certain age, ICU treatment was not offered to all patients requiring mechanical ventilation. The reason, taking into account the data in the column "estimated exitus per ICU", perhaps was to apply a policy that considered the chances of surviving an aggressive IMV intervention were limited beyond a certain age.

We made the adjustments shown in Table 3 as an attempt to estimate the number of patients dying outside the ICU who would have required admission. To do this, we estimated the number of patients dying in the ICU by multiplying the estimated value of "exitus per ICU" by the value of patients admitted to ICU (Column N). In Column O, we calculated the difference between exitus and the Column N value. In Column P we estimated the number of patients who must have actually needed mechanical ventilation (either invasive or non-invasive). The set of patients on paths R2 and R3 together will equal the value in Column P minus the patients who die (which will be path R4). This value was converted to a proportion of hospitalized patients in each age range (in Column R).

In order to differentiate how many of these patients were on path R2 and how many were on path R3, several hospitals were consulted. However, we were unable to obtain an estimate of these values (only ICU admissions, i.e., those requiring IMV, were available). Therefore, we assumed 50.0% of the patients requiring IMV were patients on path R2 and the other 50.0% were patients on path R3. This assignment was effectively arbitrary and should be informed in the future by real data (or directly eliminate path R2 by assigning 0.0% of cases when there are no NIMV-adapted facilities suitable for the care treatment of this path).

The percentage of patients on path R1 is calculated by subtracting the rest of the paths from 100.0%.

Table 3. Estimating proportion parameters for patients in each path by age

M	N	O	P	Q	R	S	T	U	v
Age range	Estimated exitus on ICU	Exitus outside ICU	Estimated number of patients that require Mechanical ventilation	R2_R3	Prop R2_R3 per hospitalized	%R1	%R2	%R3	%R4
5	6		66	60	0,04441	0,951	0,022	0,022	0,004
15	12		102	91	0,05603	0,937	0,028	0,028	0,007
25	37		228	180	0,03782	0,952	0,019	0,019	0,010
35	143		640	524	0,05517	0,933	0,028	0,028	0,012
45	514		1682	1260	0,06640	0,911	0,033	0,033	0,022

55	1587		3776	2227	0,07510	0,873	0,038	0,038	0,052
65	3437	671	6615	2507	0,06988	0,816	0,035	0,035	0,115
75	4200	4941	10221	1080	0,02738	0,741	0,014	0,014	0,232
85	935	27700	28738	103	0,00176	0,509	0,001	0,001	0,489
	=REDONDEAR.MAS (Estimated Exitus per ICU*ICU;0)	=Exitus-Estimated exitus on ICU	=ICU+Exitus outside ICU	=Estimated required Mechanical ventilation - Exitus	=R2_R3/ Hospitalized	=100%-Sum(%R2; %R3; %R4)	50%*Prop R2_R3 per hospitalized	50%*Prop R2_R3 per hospitalized	=Exit per hospitalized

The values in Columns S, T, U and V in Table 3 add up to 100%. They can be used for a simple requirement path allocation model, where only the probabilities derived from the patient age range are taken into account. In this case, patients are simply allocated randomly, according to the probabilities of the row corresponding to the age range and path.

Advanced model including comorbidity, pregnancy, and gender

It was also possible to develop a model that allowed us to make a path assignment that included, in addition to age, other risk factors such as gender, pregnancy, and comorbidities. In this case, we needed to work with odds ratios to generate an individualized risk coefficient for each subject represented by a row in the data set and that this coefficient represented all risk factors.

The odds ratio (OR) (Table 4) is the quotient of the division between the number of times something happens compared to when it does not happen when a variable is present, and the same division when the variable is not present (Dominguez-Lara, 2018; Marin-Garcia, Bonavia, et al., 2020). If the OR equals 1 it means that there is no relationship between the event and the variable because the probability of the event occurring is the same when the variable is present and when it is not.

Table 4. Odds ratio calculation

	Case (it happens)	No case (it does not happen)
Variable Present	a	b
Variable Absent	c	d
OR=(a/b) / (c/d)		

As a reference category, we consider the following: female, age range 5 years, no pregnancy, and no presence of comorbidities. Any other combination of variables in the row results in an OR with respect to the reference category that will be calculated as the multiplication of the ORs of the variables that differ from the definition of the reference category (Hair et al., 2009). Figure 5 displays a schematic of this process. Thus, for example, if one row of our data set represents a 55-year-old male, diabetic, and a smoker, the aggregate OR equals the multiplication of ORmen*OR55years*ORDiabetes*ORSmoking.

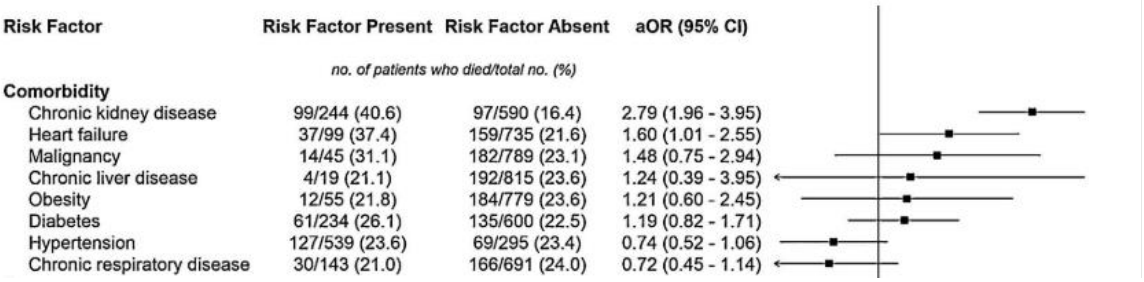
Figure 5. Odds aggregation

∀ i from 1 to num rows ; ∀ j form 1 to num of risk factors
OR_i=EXP ($b_1.x_1$+....+$b_n.x_n$)= EXP($b_1.x_1$).... EXP($b_n.x_n$)
when x_j is 0 (reference category for risk factor j) -->EXP(0)=1

The OR can be transformed into a probability by means of the formula *probability = OR/(OR+1)* (Hair et al., 2009).

We extracted OR data related to COVID-19 severity from various sources (Posso et al., 2020; Verity et al., 2020; Zheng et al., 2020). Figure 6 shows the results of the work of Posso et al. (2020). With the CNE data from the second and third waves in Valencia province, the ORs due to being male compared to being female, assuming it was constant for all age ranges, were 1.85 (10.2% ICU hospitalization for men vs. 5.6% for women). In future research, each RO can be analyzed separately for different age ranges, assuming data are available.

Figure 6. Adjusted Odds Ratios by comorbidity (Posso et al , 2020)



As we were unable to locate any work that estimated ORs by age range, we calculated them from the data provided by CNE for Spain (from September 1, 2020 to March 24, 2021). Columns a) and b) of Table 5 came from Column P of Table 3, Column D of Table 3, and Column D of Table 2. The ORs (by age) are the quotient between a/b of each range and a/b of the reference category (5 years).

Table 5. Odds ratio by age range (reference category = 5 years)

Age range	a) Estimated number of patients that require Mechanical ventilation (case)	b) Num hospitalized patients R1 (no case)	a/b	OR by age
5 (reference category)	66	1285	0,05136187	1
15	102	1522	0,06701708	1,30480229
25	228	4531	0,05032002	0,9797155
35	640	8858	0,07225107	1,40670649
45	1682	17294	0,09725917	1,89360647
55	3776	25876	0,14592673	2,84114916
65	6615	29261	0,22606883	4,40149159
75	10221	29217	0,34983058	6,81109535
85	28738	29772	0,96526938	18,7935023
Reference category (age range >5 absent) c/d=66/1285=			0,05136187	

The following two path assignment options are proposed for the advanced model:

Advanced Model Option A

In Option A, the paths are assigned taking into account the percentile which represents the value of the aggregate OR of each row with respect to the total sample generated. Higher ORs imply a higher probability of generating worse symptoms. We achieve this by assigning Path 1 when the percentile is less than or equal to the proportion of patients in R1. We assign Path 2 if the percentile is greater than the proportion of patients in R1 and less than or equal to the sum of the proportions of R1 and R2. We assign Path 3 if the percentile is greater than the sum of proportions of R1 and R2, but less than or equal to the sum of proportions R1, R2 and R3. We assign Path 4 in all other cases.

The proportions of patients in each path (R1, R2, R3, or R4) are calculated as an average weighted by the number of patients in each age range from the values in columns S, T, U, and V of Table 3.

Note the ORs in Table 5 were not adjusted for the presence of comorbidities, which tend to become more frequent with advancing age. Therefore, it is possible the data generated with this model may be implausible, showing a disproportionate presence of elderly cases in the R4 path. To address this problem, we propose option B.

Advanced Model Option B

In Option B, we start from the same ORs as in Option A. However, the percentile representing the OR of each row is calculated only by comparison with other cases within the same age range (not the entire generated sample). The assignment of paths will also be made within each age range based on that percentile and the case proportions of Columns S, T, U, and V of Table 3.

Step 4

In Step 4, each patient is assigned the date of the onset of severe and critical symptoms (thus, the patient needs to be admitted a hospital for observation or to another care unit). For this purpose, a distribution of the number of new severe cases per day is used as a starting point. This distribution can be generated from actual data. For example, in the supplementary material documents attached (https://zenodo.org/record/4699554), the number of cases generating severe symptoms in our data set tracked the daily proportion of real patients who required hospital admission in Valencia province (Spain) between September 1, 2020 and March 24, 2021.

However, other starting data sets can be used. These can be real data from other geographical areas and/or time windows. Alternately, they can be simulated data from different scenarios where a sequence of one or more waves with different intensities (peak of patients with severe symptoms on a given day) and amplitude (number of days from the start of the wave until it is considered to be eradicated) is modeled.

For the rows coinciding with each age range category and for each requirement path, the number of days the patient will have symptoms requiring observation and oxygen (DaysObs), the number of days requiring NIMV (DaysNIMV), the number of days requiring IMV (DaysIMV), and the number of days under observation after having undergone NIMV or IMV (DaysPstIMV) are generated. For this purpose, Poisson distributions with a lambda parameter equal to the expected average number of days with

symptoms for that age range and path (Petermann-Rocha et al., 2020) are used. Taking into account the duration of symptoms requiring a given treatment was shorter for people who died than for those who recovered, the patients in each path were parameterized differently. This added complexity due to the lack of real data broken down by path (Casas-Rojo et al., 2020; Guan et al., 2020; Huang et al., 2020; Rubio-Rivas et al., 2020; Wang et al., 2020; Xu et al., 2020). We used as parameters the mean values collected by a large Spanish hospital broken down by age range and final outcome (recovered or exitus). It was not yet possible to have disaggregated and different data for all the paths (so some parameters were repeated). This may cause the simulated data to have a slight deviation with respect to the evolution of the disease in real patients.

A Poisson distribution was assumed as we knew only one parameter (mean). In the case of having more information to estimate scale and shape, future research could investigate the viability of a Weibull distribution instead. (Celeux et al., 2006; Epstein & Dexter, 2020; Mun, 2008).

It was assumed the duration in days of each symptom was a memoryless process. Thus, each duration was independent of the duration of the other disease states.

Guidelines for Data Reuse

Each data set represents the demographics, comorbidities, dates of admission, and dates of transition to other states (columns) for a number of patients (rows).

The generated data can be used to test whether a model or a simulator performs with at least synthetic data. We assumed that if a model does not work with synthetic data, then it will not work with real data. The converse is not necessarily true, i.e., a model may work with synthetic data, but fail with real data.

The algorithm generates families of data sets (replications) that share the same generation parameters, but with changes to the randomization seed in each replication. The parameters set to run the algorithm are as follows:

- The number of replications to be generated (N).
- The number of patients to be generated (rows of the data set).
- Age range. Column vector with the age value representing each age range, usually the mean value of the range.
- Percentage of patients desired in each age range. Column vector with the same number of elements as age ranges considered. The sum of its values should equal 100.0%.
- Incidence of comorbidity. Percentage of patients who will have each of the considered risk factors present.
- OR of death by COVID-19 corresponding to the presence of each risk factor with respect to the reference category (female, age range 5 years, not pregnant, and without comorbidities).
- Distribution of patients by requirement paths. One column vector for each path considered. Each vector has as many elements as age ranges considered. The sum of the elements of the same row should be 100.0%.

- Percentage of cases per day with respect to the total number of cases generated. Column vector with as many elements as days contained in the time window to be generated. Real numbers of cases per day can be entered (the generator converts them into percentages), or the percentages can be entered directly. You can set a constant input each day or the data can be adjusted to the desired input curve (varying peaks, amplitude, and waves).
- The average duration values for each symptom, for each age range, and for each path. A column vector for each combination of paths and care requirements. In the example shown, with 4 paths (R1 to R4) and 4 requirements (Observation, NIMV, IMV, Observation post mechanical ventilation), 16 vectors are needed. Each vector has as many elements as age ranges considered. Paths where certain symptoms do not occur will have a duration of zero in all elements of their vector. If values disaggregated by age and/or path are not available, the same average duration value can be input for all elements.

One data set (or partial data set) can be used to calibrate a model. Subsequently, the calibrated model can be validated using the other part of the data set (or other data sets from the same or different families). In this way, the model can be checked for proper fit to a new and unused data set to refine or calibrate the model. If real data become available later, the performance of the model in a real environment can be evaluated.

For example, a family can be generated with N = 30 data sets, each with 1500 rows. You can build a model using the first 1000 rows of this data set and then test with the remaining 500 rows. Alternately, you can build a model with the complete data set and test it with the 29 family data sets or with data sets from other families.

You can create as many data set families as you want so there are variations in requirement duration times (e.g., testing scenarios where they are totally random, they follow a statistical distribution, or they are a function of comorbidities or other data set parameters).

The data generated were based on patient needs paths. Depending on available resources or other health policies, patients may or may not receive the resource they need and this has clear consequences on the outcome. For example, if an 85-year-old patient has acute respiratory distress syndrome (ARDS), they would need mechanical ventilation to survive, and that is what our data generator will identify. Additionally, this patient may be dispensed with this indicated treatment, or it may be decided that it is not worthwhile and they will be transferred to palliative care. For this reason, the data generated by our algorithm does not necessarily match the real data of patients admitted to ICU (not all patients who need mechanical ventilation actually receive it).

Data Set Description

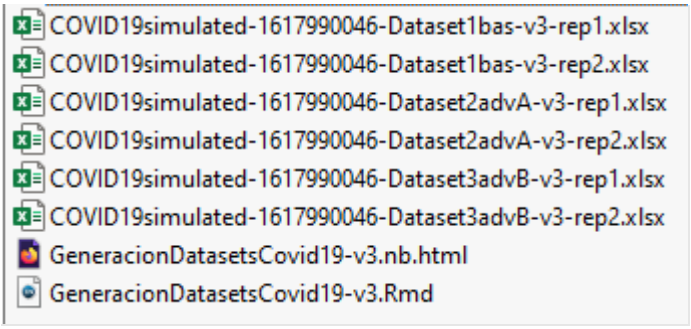
The ZIP archive provided as supplementary material (Figure 7) contains the R code in a notebook (RMD file extension) and its HTML version. In addition, the archive includes the XLSX files for two replications generated with the same parameters. Each replication consists of three data sets, one for the data generated with the basic model, another with the advanced model Option A, and the third with the

advanced model Option B. Thus, in each replication three data sets were generated (data set1bas, data set2advA and data set3advB). The data sets of each replication share parameters and randomization seeds, but assign the paths differently. Model 1 does not take into account comorbidity parameters or gender and assigns the path only according to age range so the number of cases resembles the proportion of patients in each path stipulated by the parameters. Model 2 is based on the combined risk of comorbidity, gender, and age and assigns paths based on those alone (ignoring the parameter specification of how many patients of each age range are desired in each path). Model 3 assigns paths taking into account the ORs of all risk factors (comorbidity, gender, and age range), but sorts by age before allocating paths based on risk. In this way, Model 3 also respects the specifications of the distribution parameter of patients per path according to age range.

The most appropriate model depends on the richness and reliability of the data entered as parameters. If there is little information available on the incidence and/or the OR of the risk factors, but the incidence by age range is plotted, then Model 1 would likely produce the results closest to reality. If incidence information by age range is unavailable, but risk factor ORs are reliable, then Model 2 would likely be the most appropriate version. When complete data are available, Model 3 is likely a better representation of reality.

Each replication (rep) is labeled by the number preceding the file extension. Each generation model is labeled with a number following the word "Data set" (1 for basic, 2 for Option A and 3 for Option B). The algorithm version is labeled with the number following the "V". The most current version is version 3.

Figure 7. Files included in supplementary material



The notebook file is structured in three sections including declaration of the required libraries, definition of the parameters (Figure 8), and the data generation algorithm based on the parameters (Figure 9).

Figure 8. Code Example for Parameter Specification

```
N=2 #number of random replications with the same parameters

#Simulation limits
NumRow<-13781 #number of rows in the simulated dataset (with N>10000 distributions meet populations' parameters. Useful for test)
DiaIni = as.Date("2020-09-01") #Start date for simulation.


#1 sociodemographic
#gender
MenProp<- 0.5454 #source CNE -Centro Nacional de Epidemiología. (2020a). Información científico-técnica, enfermedad por coronavirus, COVID-19 (actualizado 20201112). https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/ITCoronavirus/home.htm

#AGE
# Proportion of patients in Hospitalized bed
# source: Situación de COVID-19 en España a 22 de diciembre de 2020. Equipo COVID-19. RENAVE. CNE. CNM (ISCIII). Table 4 page 8
IdforAge<-c(5,15,25,35,45,55,65,75,85) #class interval for each of the age categories.Last category considering 90 as maximum age to simulate
AgeProp<- rep (0,length(IdforAge)) # 9 age categories
AgeProp[1]<-0.0044 #0-9 years
AgeProp[2]<-0.0057#10-19 years
AgeProp[3]<-0.0214#20-29 years
AgeProp[4]<-0.0465#30-39 years
AgeProp[5]<-0.1041#40-49 years
AgeProp[6]<-0.1529 #50-59 years
AgeProp[7]<-0.1800 #60-69 years
AgeProp[8]<-0.2019 #70-79 years
AgeProp[9]<-1-sum(AgeProp[1:length(IdforAge)-1]) #80+years


#2 comorbidity
# Source: CNE -Centro Nacional de Epidemiología. (2020a). Información científico-técnica, enfermedad por coronavirus, COVID-19 (actualizado 20201112). https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/ITCoronavirus/home.htm
#prevalence in covid hospitalized patients
CardioProp<-0.285 #heart conditions, such as heart failure, coronary artery disease, or cardiomyopathies
COPDProp<-0.116 # chronic obstructive pulmonary disease
DiabetesProp<-0.177 # Type 2 diabetes mellitus
HTAProp<-0.129 # Hypertension
KidneyProp<-0.0262 # Chronic kidney disease
CancerProp<-0.0329 # cancer
ObesityProp<-0.169 # (BMI> 30 kg/m2)
PregnacyProp<-0 #not included at present
SmokingProp<-0.23

#3 Sequence of symptoms
PathsNames<-c("R1","R2","R3","R4")

##R2 severe symptoms, observations + NIMV + Observation and recovered
percR2ByAge<-c(
0.022,
0.028,
0.019,
0.028,
```

Figure 9. Code Example for Data Set Generation Algorithm

```
#Code by steps
```{r}
# for Exporting results each Set of datasets different time run
a<-Sys.time()

#Random replications with the same parameters
seed <-458946 #random number for reproducibility
for (j in 1:N){
seed <-seed+j
set.seed(seed)

Id<-c(1:NumRow)

#1 sociodemographic
#Gender
# Create uniform random sample with MenProp men ## legend men:0; women:1
Gender<- sample(0:1, size=NumRow, prob=c(MenProp,1-MenProp), replace=TRUE)


#AGE
AgeRange<- sample(IdforAge, size=NumRow, prob=AgeProp, replace=TRUE)

#2 COMORBIDITY
Cardio<- sample(0:1, size=NumRow, prob=c(1-CardioProp,CardioProp), replace=TRUE)
COPD<- sample(0:1, size=NumRow, prob=c(1-COPDProp,COPDProp), replace=TRUE)
Diabetes<- sample(0:1, size=NumRow, prob=c(1-DiabetesProp,DiabetesProp), replace=TRUE)
HTA<- sample(0:1, size=NumRow, prob=c(1-HTAProp,HTAProp), replace=TRUE)
Kidney<- sample(0:1, size=NumRow, prob=c(1-KidneyProp,KidneyProp), replace=TRUE)
Cancer<- sample(0:1, size=NumRow, prob=c(1-CancerProp,CancerProp), replace=TRUE)
Obesity<- sample(0:1, size=NumRow, prob=c(1-ObesityProp,ObesityProp), replace=TRUE)
Pregnacy<- sample(0:1, size=NumRow, prob=c(1-PregnacyProp,PregnacyProp), replace=TRUE)
Smoking<- sample(0:1, size=NumRow, prob=c(1-SmokingProp,SmokingProp), replace=TRUE)

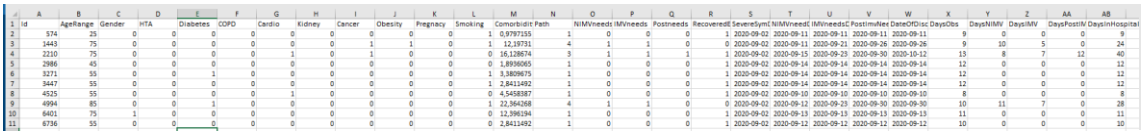


#3 Sequence of symptoms
## creation of empity data frame
SevereSymDate<-rep(0,NumRow)
ComorbidityIndex<-rep(0,NumRow)
NIMVneedDate<-rep(0,NumRow)
IMVneedsDate<-rep(0,NumRow)
PostImvNeedsDate<-rep(0,NumRow)
DateOfDischarge<-rep(0,NumRow)
RecoveredExitus<-rep(1,NumRow) #1 recovered 0 exitus (by default all recovered)
ComorbidityIndex<-rep(1,NumRow)
Path<-rep(1,NumRow) #default requirements path is R1

NIMVneeds<-rep(0,NumRow)
IMVneeds<-rep(0,NumRow)
Postneeds<-rep(0,NumRow)
```

Figure 10 shows the column headings and the first rows of the first of the generated data sets (replication 1, basic model). The description of each of the column variables is presented in Table 6. Each data set was generated with the rows sorted by date of occurrence of severe symptoms.

Figure 10. First rows of file covid-19simulated-1617990046-Dataset1-v3-1.xlsx



Id	AgeRange	Gender	HTA	Diabetes	COPD	Cardio	Kidney	Cancer	Obesity	Pregnacy	Smoking	Comorbidit	Path	NIMVneeds	IMVneeds	Postneeds	Recovered	SevereSym	NIMVneed	IMVneeds	PostImvNee	DateOfDisc	DaysObs	DaysNIMV	DaysIMV	DaysPostIM	DaysInHospital
574	25	0	0	0	0	0	0	0	0	0	1	0,9797155	1	0	0	0	1	2020-09-02	2020-09-11	2020-09-11	2020-09-11	2020-09-11	9	0	0	0	9
1443	75	0	0	0	0	0	1	1	0	0	1	12,19731	4	1	1	0	0	2020-09-02	2020-09-11	2020-09-21	2020-09-26	2020-09-26	9	10	5	0	24
2210	75	0	0	0	0	1	0	1	0	0	0	16,128674	3	1	1	1	1	2020-09-02	2020-09-15	2020-09-23	2020-09-30	2020-10-12	13	8	7	12	40
2986	45	0	0	0	0	0	0	0	0	0	0	1,8936065	1	0	0	0	1	2020-09-02	2020-09-14	2020-09-14	2020-09-14	2020-09-14	12	0	0	0	12
3271	55	0	0	1	0	0	0	0	0	0	1	3,3809675	1	0	0	0	1	2020-09-02	2020-09-14	2020-09-14	2020-09-14	2020-09-14	12	0	0	0	12
3447	55	0	0	0	0	0	0	0	0	0	1	2,8411492	1	0	0	0	1	2020-09-02	2020-09-14	2020-09-14	2020-09-14	2020-09-14	12	0	0	0	12
4525	55	0	0	0	0	1	0	0	0	0	0	4,5458387	1	0	0	0	1	2020-09-02	2020-09-10	2020-09-10	2020-09-10	2020-09-10	8	0	0	0	8
4994	85	0	0	1	0	0	0	0	0	0	1	22,364268	4	1	1	0	0	2020-09-02	2020-09-12	2020-09-23	2020-09-30	2020-09-30	10	11	7	0	28
6401	75	1	0	0	0	0	0	0	0	0	0	12,396194	1	0	0	0	1	2020-09-02	2020-09-13	2020-09-13	2020-09-13	2020-09-13	11	0	0	0	11
6736	55	0	0	0	0	0	0	0	0	0	0	2,8411492	1	0	0	0	1	2020-09-02	2020-09-12	2020-09-12	2020-09-12	2020-09-12	10	0	0	0	10

For the purpose of simplicity, we assume all patients cross all requirements (observation, NIMV, IMV, observation post Mechanical Ventilation and discharge) but the duration of unnecessary requirements is effectively zero, i.e., the day showing one requirement equals the day for the next requirement. In this way, when we make calculations using date columns, all the rows have a day in each column. Even if a
```

patient does not present IMV needs or post-IMV needs (in those cases the date of discharge is the same as the day of IMV needs or of post-IMV needs making the length of stay in those processes equal to zero). The same is true for exitus patients where the post-IMV needs day is the same as the discharge day.

Table 6. Data description

Column name	Description	Values
Id	Sequential number that uniquely identifies each row of the data set	1 to number of rows
AgeRange	Age range. Representative value of the age category to which the patient represented belongs in each row. Each age category contains 10 years (minus the last one) and the age range value represents the average value of the category (minus the last one, which is set to 85).	Integer from 5 to 85. distanciados de 10 en 10
Gender	Gender of the patient in row	0=men 1=women
HTA	Hypertension	0=absent 1=present
Diabetes	Type 2 diabetes mellitus	0=absent 1=present
COPD	chronic obstructive pulmonary disease	0=absent 1=present
Cardio	heart conditions, such as heart failure, coronary artery disease, or cardiomyopathies	0=absent 1=present
Kidney	chronic kidney disease	0=absent 1=present
Cancer	cancer	0=absent 1=present
Obesity	obesity (BMI> 30 kg/m2	0=absent 1=present
Pregnacy	pregnancy	0=absent 1=present
Smoking	smoking	0=absent 1=present
ComorbidityIndex	Comorbidity index. Multiplication of the Odds Ratio of risk factors (age range, genre, comorbidity and pregnancy)	Real number >0
Path	Path requirement of the patient in row	Integer 1 to 4
NIMVneeds	NIMV required by patient in row (depends on path value)	0=no 1=yes
IMVneeds	IMV required by patient in row (depends on path value)	0=no 1=yes
Postneeds	Observation bed after NIMV or IMV required by patient in row (depends on path value)	0=no 1=yes
RecoveredExitus	Outcome of patient in row (depends on path value)	0=exitus 1=recovered
SevereSymDate	Calendar day when severe symptoms appear and admission in hospital was needed	Date (YYYY-MM-DD)
NIMVneedDate	Calendar day when critical symptoms appear and non invasive mechanical ventilation was needed. Calculated based on days of symptoms and date of other stages in the process	Date (YYYY-MM-DD)
IMVneedsDate	Calendar day when needs for invasive mechanical ventilation appear. Calculated based on days of symptoms and date of other stages in the process	Date (YYYY-MM-DD)
PostImvNeedsDate	Calendar day when patient does not need invasive mechanical ventilation and only need observation in hospital bed. Calculated based on days of symptoms and date of other stages in the process	Date (YYYY-MM-DD)
DateOfDischarge	Calendar day for discharge (recovered or exitus). Calculated based on days of symptoms and date of other stages in the process	Date (YYYY-MM-DD)
DaysObs	Number of days with severe symptoms in observation	Integer >0
DaysNIMV	Number of days with critical symptoms that require NIMV	Integer >0
DaysIMV	Number of days with critical symptoms that require IMV	Integer >0
DaysPostIMV	Number of days with severe symptoms in observation after MV	Integer >0
DaysInHospital	sum of all time with severe or critical symptoms (that is, the difference between admission date and discharge date)	Integer >0

Future Lines of Research Using this Data Set

This paper opens up potential future lines of research, both theoretical and practical. From a theoretical point of view, it would be interesting to develop machine learning tools that, by analyzing specific data samples from real hospitals, would be able to identify the necessary parameters for the automatic prototyping of generators adapted to each hospital. Regarding applied lines of research, clearly the proposed formalism for the generation of plausible patients is not limited to patients affected by SARS-CoV-2 infection. The generation of heterogeneous patients can serve to represent the needs of a specific population and serve as a basis for studying the behavior of complex health care delivery systems.

These data sets can be useful for earlier stages of data modeling and for testing the feasibility of discrete event models when actual data are not yet available (e.g., awaiting ethics committee acceptances, pending collection and processing by health facilities).

Actual data provided by hospitals are often incomplete or inconsistent. Therefore, they are not necessarily useful for optimization validation or simulation models. The framework presented can serve as a starting point for researchers to extend or replicate the model.

As the COVID-19 information and trace systems evolve, the plausibility of these data sets could be validated and possible predictors for paths not yet considered could be analyzed.

For those in hospital management positions, one advantage of this research is they do not have to start from scratch. It is foreseeable that the emergence of COVID-19 variants, the effect of vaccination campaigns, and new respiratory-assisted technologies will change the percentages of patients travelling along each path. These changes can be easily modeled by adjusting the parameters to quickly produce a data set that represents the new situation. In addition, our approach allows managers to identify what information to collect, store, and process in their information systems to better support COVID-19 bed management decisions.

Our data sets can also be used to estimate patient needs and differentiate them from data reflecting hospital bed management policies. In this way, they can be used to analyze scenarios and assess the effects of certain alternative policies, or to analyze the sensitivity of bed management decisions to deviations from baseline parameters.

Additionally, during the parameterization phase of the generator process, it has to be supplied with data that may not be available and are perhaps appropriate for future research, e.g., the effect of comorbidities adjusted for gender and age, or the incidence of simultaneous comorbidities.

It would also be interesting to verify to what extent and under what circumstances each of the data generation models provides more plausible results.

Similarly, the model could be extended by incorporating more trajectories or more stages of the process (e.g., from mild or asymptomatic infections and home care requirements).

Another line of future research might be concerned with identifying parameter values to feed into the models. For example, studying the incidence of comorbidities in certain populations or the incidence of several comorbidities simultaneously in the same person. Another interesting improvement would be the estimation of ORs with respect to the reference category, of needing NIMV, IMV, or death adjusted for

gender, age, and other comorbidities. It is likely the ORs are dependent on each other. That is, the risk of suffering certain comorbidities is associated with age or gender. The data available to date are not usually adjusted for other potentially confounding variables.

On the other hand, this data set can be used to check the effect of various hospital management decisions that contribute to improving the experience of the patient or healthcare staff. For example, better estimates of bed occupancy can help improve decision efficiency in ancillary services such as laundry, warehouses, kitchens; improve the warehouse management or purchases of sanitary supplies or protective equipment; or improve the efficiency of allocating health workers shifts.

Finally, this generator could be adapted to other diseases besides COVID-19 which lead to a high consumption of health care resources, where their trajectories are well studied and their parameters can be easily obtained.

Access to the data set

The data sets, called "COVID19simulated-#######-Dataset#-v3-#.xlsx" in file MarinEtAl(2021)COVID19simulated-DOI10.4995wpom.15332.zip, are deposited in https://zenodo.org/record/4699554

Other useful information related to covid-19

- https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/situacionActual.htm . The disadvantage of this page is that it only provides recent data. It does not offer historical downloads that could be used to analyze trends
- Situación de COVID-19 en España
- Actualización nº331: enfermedad por SARS-CoV-2 (COVID-19) 12.03.2021
- Informe de indicadores principales de seguimiento de COVID-19 (actualización semanal) 11.03.2021
- http://coronavirus.san.gva.es/es/estadisticas
- https://dadesobertes.gva.es/va/dataset?tags=COVID-19
- https://cnecovid.isciii.es/covid19/#documentaci%C3%B3n-y-datos
- casos_hosp_uci_def_sexo_edad_provres.csv

Note many of these information sources provide data in formats of low usability.

Author contributions

All authors listed have made a substantial, direct and intellectual contribution to the work and approved it for publication.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Acknowledgments

This research received no external funding. This research was funded by personal funds from the authors and support from the ROGLE research group.

Spanish version

Resumen

En este trabajo se presenta cómo se ha generado un conjunto de datos verosímiles relacionados con las necesidades de pacientes covid-19 con síntomas severos o críticos. Se considerarán las etapas posibles con los conocimientos médicos a fecha de enero de 2021. Los parámetros elegidos en este data set *están personalizados para adecuarse a los valores poblacionales de la región de Valencia (España), unos 2.5 Millones de habitantes y la evolución de la pandemia entre los meses de septiembre 2020 y marzo 2021, un periodo de tiempo que contempla dos olas completas de pandemia.*

En contra de lo que cabría esperar, a pesar de la ley de transparencia europea y nacional (BOE-A-2013-12887, 2013; Parlamento Europeo y del Consejo de la Unión Europea, 2019), los datos reales relacionados con la pandemia covid-19, al menos en España, tardan mucho en actualizarse y estar disponibles (normalmente una semana o más días). Además, algunos datos relevantes para trabajar los modelos de gestión de camas de hospital no están accesibles públicamente. Bien porque no se hayan recogido esos datos, o porque los organismos públicos no los ofrecen (a pesar de tenerlos indexados en sus bases de datos), o los ofrecen camuflados en indicadores procesados y no muestran los datos en bruto, o simplemente los publican en un formato de difícil reutilización (por ejemplo, en documentos PDF en lugar de en tablas CSV). A pesar de que los sistemas de información de los hospitales son bastante potentes, siguen existiendo datos que ni siquiera están recogidos adecuadamente en el sistema de información de salud.

Por otra parte, los datos recogidos en un hospital dependen de las estrategias y prácticas propias de ese hospital o sistema de salud. Este efecto limita la generalización de los datos "reales" y es necesario trabajar con datos "realistas" o verosímiles que están limpios de interacciones con variables o decisiones locales (Gunal, 2012; Marin-Garcia et al., 2020).

Por un lado, se puede parametrizar el modelo y definir la estructura de datos que sería necesaria para ejecutar el modelo con datos reales. Por otro lado, se pueden generar conjuntos de datos verosímiles a partir de la información pública disponible y, posteriormente, cuando se disponga de los datos reales evaluar la bondad del modelo (Garcia-Sabater & Maheut, 2021).

Este trabajo abre líneas de investigación futura tanto teóricas como prácticas. Desde el punto de vista teórico, sería interesante el desarrollo de herramientas de aprendizaje automático que, analizando muestras de datos especificas en hospitales reales, sean capaces de identificar los parámetros necesarios para el prototipado automático de generadores adaptados a cada hospital. En cuanto a las líneas de investigación aplicadas, es evidente que el formalismo propuesto para la generación de pacientes verosímiles no se limita a pacientes afectados por la infección del SARS-CoV-2. La generación de pacientes heterogéneos puede servir a representar las necesidades de una población específica y servir de base para el estudio del comportamiento de sistemas de prestación de servicios de salud complejos.

Introducción

Desde los comienzos de la crisis sanitaria asociada a la covid-19, cada sistema sanitario ha tenido que enfrentarse a requerimientos de atención sanitaria fluctuantes, haciendo frente a olas de distintas amplitudes y duraciones, condicionadas en parte por la aplicación de restricciones y protocolos sanitarios. Desde el principio de la crisis, muchos científicos han trabajado en diseñar soluciones para mitigar los efectos de la covid-19. Los ámbitos han sido múltiples, desde el desarrollo de vacunas, hasta herramientas de previsión de infecciones y del impacto de medidas políticas de mitigación utilizando las técnicas más avanzadas de inteligencia artificial. Sin embargo, la gestión de hospitales y de sus recursos no parece haber recibido tanta atención (Epstein & Dexter, 2020). Se puede observar que cada país, cada región y hasta cada hospital ha llegado a gestionar sus recursos críticos (camas y personal sanitario) de manera local, sin que existan mecanismos de coordinación y herramientas para anticipar y mitigar las consecuencias de las olas, que normalmente se producen por agregación de clústeres locales de infecciones.

La gestión de camas de hospitales es una aplicación concreta de un problema genérico de gestión de la capacidad (Claudio et al., 2021; Garcia-Sabater et al., 2020; Lagarda-Leyva & Ruiz, 2019; Marin-Garcia et al., 2019; Nino et al., 2021; Xia & Sun, 2013). Para ello, es factible el intento de usar las herramientas de gestión de operaciones, y más concretamente las herramientas de diseño, planificación y control o mejora de procesos. En este sentido, la simulación basada en eventos discretos es una herramientas para dar soporte a las toma de decisiones de gestión hospitalaria (Gunal, 2012; Marin-Garcia, Garcia-Sabater, et al., 2020). Con un simulador de procesos, se podría facilitar una planificación adecuada de los recursos asistenciales y anticipar, o al menos mitigar, situaciones donde en unos centros de salud no se pueda atender a pacientes debido a la saturación o colapso del sistema, mientas que en otros centros haya recursos ociosos (Romeo Casabona & Urruela Mora, 2020). También podrían servir para determinar empíricamente umbrales de ocupación antes de desviar pacientes entre hospitales y, de este modo, evitar

transferir pacientes en estados más avanzados de la enfermedad y, por tanto, con un traslado mucho más complicado, arriesgado y costoso.

Hemos de tener en cuenta que, en esta aproximación al problema de gestión de camas de hospital, no se intenta predecir si una persona concreta se recuperará, pasará a un estado de mayor gravedad o morirá; ni tampoco predecir los días exactos que estará un-a paciente concreto en cada etapa. El objetivo es modelar los cuidados necesarios para la atención de pacientes covid-19 y los procesos asociados dentro del hospital, con el fin de predecir, con suficiente fiabilidad, la ocupación global diaria de camas de hospitalización, el uso de equipamiento de ventilación mecánica no invasiva (NIMV), la ocupación de camas en unidades de cuidados intensivos (ICU), la necesidad de personal médico, así como la necesidad de derivación de pacientes dentro del área de estudio elegida. En este sentido, conviene aclarar que la predicción consistiría en estimar probabilidades para diferentes tasas de ocupación en determinados días del futuro o, dada una ocupación actual y teniendo en cuenta la tasa de ingresos covid-19 en el hospital, cuándo dejará de disponerse de camas libres en hospitalización o ICU y sería necesario derivar pacientes o habilitar más capacidad para el circuito de pacientes covid-19.

Uno de los problemas a resolver es que el modelado de las distribuciones estadísticas, tanto de los tiempos de requerimiento de cuidados, como de las diferentes etapas que puede atravesar un paciente covid-19 requiere de un volumen de datos importante (Gunal, 2012). En contra de lo que cabría esperar, a pesar de la ley de transparencia europea y nacional (BOE-A-2013-12887, 2013; Parlamento Europeo y del Consejo de la Unión Europea, 2019), los datos reales relacionados con la pandemia covid-19, al menos en España, tardan mucho (normalmente una semana o diez días) en actualizarse y estar disponibles. Además, algunos datos relevantes para trabajar con los modelos de gestión de camas de hospital no están accesibles públicamente. Esto puede tener diferentes orígenes: que no se haya recogido esos datos, los organismos públicos no los ofrecen (a pesar de tenerlos indexados en sus bases de datos internas), que los organismos públicos los ofrezcan "camuflados" en indicadores procesados y no muestren los datos en bruto; o simplemente que los publiquen en un formato de difícil reutilización (por ejemplo, en documentos PDF en lugar de en tablas CSV).

Por otra parte, a pesar de que los sistemas de información de los hospitales son bastante potentes, siguen existiendo datos que ni siquiera están recogidos adecuadamente en el sistema de información de salud pública. Son datos que no se registran en el proceso de admisión o tratamiento de los pacientes o, en el caso de registrarse, se hacen en campos o con formato no estandarizado. Esto obliga a una extracción y filtrado manual de la información que puede generar errores e impide un volcado y acceso sencillo a dichos datos.

Otro problema, aunque no el menor, reside en la dependencia que existe entre los datos recogidos en un hospital y las estrategias y prácticas propias de ese hospital o sistema de salud. Así, cuando se observan los datos, en realidad lo que se observa es el resultado de una serie de políticas aplicadas. Algunas de esas políticas son explícitas. Pero otras son implícitas y, en muchos casos, no observables. Este efecto limita la generalización de los datos "reales" y, en muchas ocasiones, es mucho más interesante para los investigadores trabajar con datos "realistas" o verosímiles que están limpios de interacciones con variables o decisiones locales.

Por estas razones, para la simulación de procesos internos del tratamiento de pacientes covid-19 en hospitales, puede ser adecuado hacerlo con datos realistas en vez de reales (Gunal, 2012). Sería un error

retrasar la creación del modelo a la espera de los datos reales (Garcia-Sabater & Maheut, 2021). Por un lado, se puede parametrizar el modelo y definir la estructura de datos que sería necesaria para ejecutar el modelo sin que se disponga de los datos reales. Por otro lado, se pueden generar conjuntos de datos verosímiles a partir de la información pública disponible y, posteriormente, cuando se disponga de los datos reales evaluar la bondad del modelo (Garcia-Sabater & Maheut, 2021).

Para intentar solventar esta situación, proponemos en este trabajo un algoritmo para la generación de datos verosímiles relacionados con las necesidades de atención sanitaria de pacientes de covid-19 sintomáticos (con síntomas severos o críticos). Utilizaremos este algoritmo para generar un data set que sirva de banco de pruebas para los modelos de simulación que se generen en el futuro. De este modo, se podrá comprobar el efecto de distintas decisiones relacionadas con la gestión de camas de hospital (niveles de triaje o descarga de pacientes, derivaciones, aumento de capacidad, etc.), en las trayectorias y resultados de los pacientes generados. Además, este *data set* permitirá comprobar si, para la generación de modelos de simulación de eventos discretos para la gestión de camas de hospital, es suficiente con las tablas de datos generadas o si son necesarias variables adicionales no contempladas aún en los data set propuestos.

El conjunto de datos generado en este trabajo facilita la rápida generación de nuevas investigaciones, la reproducibilidad de las mismas y la validación de los resultados (Marin-Garcia, 2015; Roa-Martínez et al., 2017). Gracias a la reutilización de los datos o la creación de nuevos conjuntos por medio del script que proporcionamos se podrán comparar la bondad de diferentes modelos o surgir nuevas conclusiones analizando el mismo conjunto de datos, pero utilizando técnicas o enfoques alternativos.

Metodología

En la generación de datos se ha utilizado el programa R-Studio (RStudio Team, 2020) usando diversos paquetes R (Comtois, 2021; R Core Team, 2020; Revelle, 2021; Ruckdeschel et al., 2006; Schauberger & Walker, 2020; Venables & Ripley, 2002; Wickham, 2007, 2011; Wu et al., 2020): MASS, summarytools, stats, psych, plyr,dplyr, distr, ExtDist, openxlsx, reshape2.

Declaración de ética

No hay tratamiento de datos personales, ni solicitud de datos para participantes humanos. De acuerdo con la legislación local y los requisitos institucionales no se requirió revisión ni aprobación de comités de éticas para esta investigación.

Objetivo

El objetivo fundamental del método propuesto es crear un conjunto de datos que represente las necesidades clínicas de los pacientes covid-19 ingresados en hospitales en un área geográfica, en un periodo de tiempo que contemple dos olas completas de pandemia. Se considerarán las etapas posibles

con los conocimientos médicos a fecha de enero de 2021. Si en el futuro se incorporan nuevas etapas de tratamiento o se considera adecuado eliminar alguna de las contempladas en el modelo propuesto, se podrá modificar el método para ajustarlo a la nueva realidad.

Los parámetros elegidos en este *data paper* están personalizados para adecuarse a los valores poblacionales de la región de Valencia (España), que cuenta con unos 2.5 Millones de habitantes, representando la evolución de la pandemia entre los meses de septiembre 2020 y marzo 2021. En cualquier caso, el horizonte temporal cubierto, la intensidad de número de ingresos hospitalarios en el periodo y la curva de incidencia (más picuda o más plana) son totalmente personalizables para ajustarse a cualquier otro escenario.

Supuestos previos para la generación de datos

Este trabajo se va a centrar en los flujos de necesidades de los pacientes. Es decir, la secuencia de evolución de síntomas que hace que un paciente tenga determinadas necesidades de atención médica. Por lo tanto, no vamos a abordar las trayectorias (la programación, coordinación, interacción y asignación de recursos de todos lo pasos asistenciales necesarios dentro de un centro de salud (Alexander, 2007; Corbin & Strauss, 1988; Pinaire et al., 2017; Unroe et al., 2010)), ya que las trayectorias están condicionadas no solo por las necesidades de los pacientes sino también por la disponibilidad de los recursos.

El algoritmo de generación de datos modeliza la necesidad de recursos clínicos que tendría el paciente, no los recursos efectivamente asignados al paciente tras aplicar una política de triaje particular en un hospital. Para generar los datos de pacientes, nos hemos basado en la información facilitada por médicos internistas, intensivistas, directores de área médica y la información disponible públicamente a finales de diciembre 2020, que resumimos a continuación:

1. No existe un tratamiento que permita cambiar el curso de las etapas que atravesará un paciente (Plaza, 2021). Es decir, solo existe un tratamiento específico para la covid-19 (facilitar oxígeno), que se aplica a los pacientes que lo necesiten. El oxígeno se puede aplicar en tres grados de intensidad: oxigenación (mascarilla en cama normal de hospitalización), ventilación forzada no invasiva (NIMV por sus siglas en inglés), ventilación forzada invasiva (IMV por sus siglas en inglés) (Daniel et al., 2021; European center for disease prevention and control, 2020; Fowler et al., 2020; Manninen, 2020; Marin-Garcia, Garcia-Sabater, et al., 2020; Winck & Scala, 2021)
2. La exposición al virus no garantiza siempre infección. Pero, una vez infectado (expuesto a una dosis suficiente para desarrollar la "enfermedad) cada persona desarrolla una trayectoria que está predeterminada (aunque es desconocida) en el momento de la infección. La trayectoria de los pacientes es un conjunto de etapas, en un orden predefinido, con unos tiempos de duración variable en cada una de ellas (pudiendo ser cero el tiempo en alguna) (Fowler et al., 2020; Wong et al., 2020)
3. La enfermedad tiene un proceso progresivo y de velocidad variable (aunque en la mayoría de los casos es lento. La evolución de un estado a otro suele tardar varios días). Ningún paciente necesita un ingreso en ICU sin antes haber necesitado ingresar en un hospital para observación (Belciug et al., 2020; Olivieri et al., 2021; Stang et al., 2020), y tampoco debería ser dado de alta tras ICU sin pasar por un periodo de observación en camas de hospitalización (Castelnuovo et al., 2020; ECDP, 2020)

4. Cada fila representará a un-a paciente. Para la distribución de género, edad y comorbilidad no se tendrá en cuenta los valores de las otras celdas de la fila del-la paciente (en versiones futuras se puede mejorar el modelo representando la relación entre las variables).

Si alguno de estos supuestos no es correcto en algún contexto determinado, se pueden añadir parámetros o restricciones en el futuro para representar más adecuadamente una situación concreta.

Flujos de necesidades de pacientes

En la Figure 1 se muestra de manera esquemática la evolución de síntomas de los pacientes y el tipo de necesidades que requiere en cada momento (Daniel et al., 2021; Fowler et al., 2020; Winck & Scala, 2021). De las personas infectadas por SARS-CoV-2 un pequeño porcentaje desarrollará síntomas de covid-19 severo o crítico. En condiciones ideales, estas personas requerirán observación médica específica, normalmente con ingreso en un hospital para su seguimiento y cuidado. La dotación habitual de todas las camas de hospital en la sanidad pública española permite aplicar terapia de oxígeno sin ventilación forzada en el caso de que fuese preciso. Algunos de los pacientes mejorarán al cabo de unos días y podrán completar su recuperación en su domicilio, cuando sus síntomas sean moderados o suaves (R1). Sin embargo, otros empeorarán y requerirán de ventiladores forzados no invasivos (NIMV) (Brochard, 2003; Popat & Jones, 2012; Ruza, 2008; Wiersema, 2007) que se pueden aplicar en habitaciones normales a las que se les ha hecho una pequeña adaptación de bajo coste. Estas adaptaciones permitirían aliviar la presión de camas ICU pues los pacientes no precisan todavía de las condiciones especiales de una ICU. Tras un periodo de tratamiento con ventilación forzada, una parte de los pacientes mejorará y podrá regresar a una cama de hospital normal, donde continuará su recuperación antes de ser dados de alta y pasar a atención domiciliaria (R2). El resto tendrá síntomas más graves y necesitará de ventilación forzada invasiva (IMV) (Brochard, 2003; Buckley & Gillham, 2007; Popat & Jones, 2012; Ruza, 2008), que solo puede administrarse en camas ICU o asimiladas (ICU pediátricas, quirófanos o ICUs de emergencia). Los pacientes que se recuperen tras la estancia en ICU pasarán a una cama de hospital (post-ICU) para completar la recuperación antes de ser dados de alta y volver a sus hogares (R3). El resto fallecerán (R4).

La Table 1 muestra el desglose del diagrama de flujo de la Figure 1 en los diferentes flujos de necesidades que se contemplaran en este trabajo. El generador solo se centrará en los pacientes que requieren hospitalización -severo y crítico- (Fowler et al., 2020). Es decir, los flujos R1, R2, R3 y R4. En cada flujo de necesidades (trayectoria) puede haber diferentes "familias de pacientes", si la duración de uso de recursos es distinta en función de algún conjunto de variables identificables. Si no, simplemente habrá una dispersión en los datos del flujo de necesidades que no se puede explicar/agrupar por familias. En este trabajo se considerarán familias de pacientes agrupados por rangos de edad.

Procedimiento

Se han utilizado cifras oficiales del Centro Nacional de Epidemiología de España (CNE -Centro Nacional de Epidemiología- https://cnecovid.isciii.es/) sobre el total de casos, pacientes hospitalizados, pacientes en ICU

y defunciones por día, provincia, grupo de edad y sexo entre 1 de enero de 2020 al 24 de marzo de 2021. Este proceso ha sido imperfecto: el conjunto de datos con casos tarda algún tiempo en actualizarse, existen posibles inconsistencias menores entre los protocolos de las Regiones y los datos de la CNE solo dan totales para grupos a lo largo del tiempo, lo que nos obliga a hacer suposiciones sobre el tiempo entre pasos en las rutas de requisitos de pacientes utilizando diferentes fuentes. Finalmente, no hay información completa sobre la comorbilidad en los datos de la CNE y hacemos suposiciones basadas en otras fuentes.

El proceso de generación de datos se realizará por partes (Figure 2).

Paso 1

En primer lugar, se crearán las variables sociodemográficas. Nuestra simulación solo se basa en pacientes graves o en estado crítico que requieren hospitalización. La distribución por sexo y edad de estos pacientes puede ser diferente de la de la población general en la provincia de Valencia (Generalitat Valenciana. Conselleria de Sanitat Universal i Salut Pública, 2019; Generalitat Valenciana, 2018). El número de hombres hospitalizados es 1.2 veces el número de mujeres hospitalizadas (hombres: 54.54%; mujeres = 45.45%) (CNE -Centro Nacional de Epidemiología, 2020). Respecto a la edad, se trabajará con los rangos de edad que maneja el CNE, siendo 0.4 % de los casos menores de 10 años; 0.6% entre 10 y 19 años; 2.4% entre 20 y 29; 4.6% entre 30 y 39 años; 10.4% entre 40 y 49 años; 15.3% entre 50 y 59 años; 18.0% entre 60 y 69 años; 20.2% entre 70 y 79; y, el resto, con 80 o más años (CNE -Centro Nacional de Epidemiología- https://cnecovid.isciii.es/).

Paso 2

En segundo lugar, la comorbilidad se distribuirá de manera aleatoria entre la muestra generada para respetar una incidencia parecida a la que hay en la muestra de pacientes COVID (CNE -Centro Nacional de Epidemiología, 2020). Los datos para España están disponibles sólo hasta el informe 32 (21 de mayo 2020), a partir de esa fecha no se informa de este dato. Tampoco hemos podido localizar información sobre la prevalencia simultánea de comorbilidades o la relación de éstas con género y edad en la población general (Posso et al., 2020). Por este motivo, se han distribuido de manera aleatoria para mantener una proporciones de casos donde la prevalencia sea de un 28.5% de afecciones cardíacas, como insuficiencia cardíaca, enfermedad de las arterias coronarias o miocardiopatías; 11.6% de la enfermedad pulmonar obstructiva crónica; 17.7% diabetes mellitus tipo 2; 12.9% de hipertensión; 2.62% enfermedad renal crónica; 3.29% cáncer. El CNE no proporciona datos de algunos fatores de riesgo (16.9% obesidad (BMI$\geq$ 30 kg/m^2); 23% fumador-a) por lo que se han extraído de otras fuentes (Ministerio De Sanidad, Servicios Sociales e Igualdad, 2017). En este trabajo no vamos a incluir embarazo como factor de riesgo (no es una comorbilidad), en el futuro se puede incluir esta variable, asignando casos solo a mujeres en rango de edad fértil. En investigación futura, también se puede mejorar la generación de casos, teniendo en cuenta las razones de odds de comorbilidad en función de las variables sociodemográficas y de padecer varias comorbilidades simultáneamente (Zheng et al., 2020). Además, los valores se pueden adaptar a otras poblaciones, incorporando información de otras fuentes, como por ejemplo https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/evidence-table.html.

Paso 3

En el tercer paso se asigna a cada paciente una trayectoria en función de los riesgos derivados de sus variables demográficas y comorbilidad. A pesar de haberlos solicitado formalmente a diversos organismos, no ha sido posible conseguir datos públicos que nos permitieran establecer el porcentaje de pacientes en cada trayectoria por rango de edad. Por ello, hemos estimado esos porcentajes a partir de los únicos datos reales disponibles (CNE -Centro Nacional de Epidemiología- https://cnecovid.isciii.es/). Hemos usado los datos de España (Figure 4 y Table 2). No hemos usado los datos de la provincia de Valencia (Figure 3) por haber muy pocos casos de pacientes menores de 30 años, lo que hacía poco fiables las estimaciones de los parámetros. Los datos del CNE no representan adecuadamente las trayectorias de personas con más de 60 años. Puede parecer, engañosamente, que la incidencia de *Severe Acute Respiratory Syndrome* (SARS) se reduce a partir de esa edad. Sin embargo, la realidad es la contraria, este evento crece de manera exponencial, pero la aplicación de protocolos de atención médica hace que la mayoría de los pacientes de edades avanzadas no reciban tratamiento en ICU aunque tengan necesidad de IMV. El motivo es que la probabilidad de sobrevivir a una intervención tan agresiva es muy escasa (siendo prácticamente cero en pacientes de muy avanzada edad).

Para poder establecer los parámetros del porcentaje de personas en las rutas R3 y R4 se ha ajustado la curva de puntos que corresponde a la proporción de pacientes *ICU per hospitalized* y de porcentaje de exitus por número de pacientes ICU, con una exponencial (Figure 4 y Table 2). Con esos valores calculados (columnas K and L in Table 2) se han ido restando del flujo de pacientes hospitalizados para calcular el peso de cada uno de las rutas que representan la evolución de síntomas.

Teniendo en cuenta el supuesto previo número tres, la ruta R4 tendrá una prevalencia igual a la proporción de personas fallecidas respecto al total de personas hospitalizadas. Estos pacientes habrán requerido ventilación mecánica (tanto si se la han dado como si no).

El conjunto de pacientes de las rutas R3 y R2 está representado por las personas que han requerido ventilación mecánica. Este dato se puede estimar a partir de las personas que han requerido ingreso en UCI (pues en España hay muy pocos hospitales que ofrecieran NIMV en los primeros 10-12 meses de la epidemia). Sin embargo, en la Table 2 se puede apreciar con claridad que la tendencia de ingresos ICU se estanca y decrece a partir del rango de edad de 65 años, mientras que el número de pacientes fallecidos (exitus) crece exponencialmente. Esto parece indicar que, a partir de una determinada edad, no se ofrece el tratamiento en ICU a todo paciente que requiere ventilación mecánica. El motivo, teniendo en cuenta los datos de la columna "*estimated exitus per ICU*", quizás haya sido el aplicar una política que considera que las probabilidades de sobrevivir a una intervención tan agresiva como una IMV son muy limitadas a determinadas edades.

Hemos realizado los ajustes mostrados en la Table 3 para intentar estimar el número de pacientes que fallecen fuera de la ICU, y que la habrían necesitado. Para ello, hemos estimado el número de pacientes que morían en la ICU multiplicando el valor estimado de "*exitus per ICU*" por el valor de pacientes ingresados en ICU (columna N). En la columna O, hemos calculado la diferencia entre exitus y el valor de la comuna N. En la columna P hemos estimado el número de pacientes que realmente deben haber necesitado ventilación mecánica (sea invasiva o no invasiva). El conjunto de pacientes de las rutasR2 y R3 juntos, será igual al valor de la columna P menos los pacientes que fallecen (que serán la ruta R4).

Este valor es convertido a una proporción sobre pacientes hospitalizados en cada rango de edad (en la columna R).

Para diferenciar cuántos de estos pacientes eran de la ruta R2 y cuántos de la ruta R3 se han realizado consultas a varios hospitales. Sin embargo, no hemos sido capaces de obtener una estimación de esos valores (solo tienen información ingresos en UCI, es decir, los que requieren *Invasive Mechanical Ventilation*). Por ello, consideraremos un 50% de los pacientes que necesitan IMV, como pacientes en la ruta R2 y el otro 50% como pacientes en la ruta R3. Esta asignación es totalmente arbitraria y debería adaptarse en el futuro a datos reales (o eliminar directamente la ruta R2, asignándoles un 0% de casos, cuando no haya instalaciones adaptadas para NIMV adecuadas para el tratamiento asistencial de esta ruta).

El porcentaje de pacientes de la ruta R1 se calculará restando el resto de rutas al 100%.

Los valores de las columnas S, T, U y V de la Table 3 suman 100%. Pueden ser usados para un modelo sencillo de asignación de rutas de requisitos de pacientes, donde solo se tiene en cuenta las probabilidades derivadas del rango de edad del paciente. En este caso, los pacientes simplemente se reparten de manera aleatoria, de acuerdo con las probabilidades de la fila correspondiente al rango de edad y ruta.

Modelo avanzado que incluye comorbilidad, embarazo y género

No obstante, también se puede elaborar un modelo que permita hacer una asignación de rutas que incluya, además de la edad, otros factores de riesgo como el género, embarazo y comorbilidad. En este caso, necesitamos trabajar con las *odds ratios* para generar un coeficiente de riesgo individualizado para cada sujeto representado por una fila en nuestro *data set* y que este coeficiente represente todos los factores de riesgo.

La *odds ratio* (OR) (Table 4) es el cociente de la división entre el número de veces en los que sucede algo respecto a que no sucede, cuando está presente una variable; y esa misma división cuando no está presente la variable (Dominguez-Lara, 2018; Marin-Garcia, Bonavia, et al., 2020). Si el OR = 1 significa que no hay relación entre el evento y la variable, pues la probabilidad de que el evento ocurra es la misma cuando está presente la variable y cuando no lo está.

Consideraremos como categoría de referencia: mujer, rango de edad 5 años, no embarazo y sin presencia de comorbilidades. Cualquier otra combinación de variables en la fila, dará lugar a una OR respecto a la categoría de referencia que se calculará como la multiplicación de las OR de las variables que difieran de la definición de la categoría de referencia (Hair et al., 2009). Este proceso está esquematizado en la Figure 5. Así, por ejemplo, si una fila de nuestro data set representa a un hombre de 55 años, diabético y fumador; la OR agregada sería igual a la multiplicación de las ORmen*OR55years*ORDiabetes*ORSmoking

La *odds ratio* se puede transformar en una probabilidad mediante la fórmula: probabilidad=OR/(OR+1) (Hair et al., 2009).

Los datos de OR relacionados con severidad de COVID los hemos extraído de diferentes fuentes (Posso et al., 2020; Verity et al., 2020; Zheng et al., 2020). En la Figure 6 mostramos los resultados del trabajo de Posso et al (2020). Con los datos de la segunda y tercera hola en la provincia de Valencia, las OR debidas a ser hombre respecto a ser mujer, asumiéndolo constante para todos los rangos de edad, son 1.85

(10.2% ICU vs Hospitalizen for men vs 5.6% for women. Source: CNE). En una investigación futura se puede analizar de manera separada cada OR para diferentes rangos de edad, si existirán datos al respecto.

Como no hemos localizado ningún trabajo que estime las *Odds ratio*s por rango de edad, los hemos calculado a partir de los datos facilitados por CNE para España (de 1 septiembre 2020 a 24 marzo 2021). Las columnas a) y b) de la Table 5 provienen de la columna P de la Table 3 y la columna D de la Table 2. Las *odds ratio* (*OR by age*) son el cociente entre a/b de cada rango y el a/b de la categoría de referencia (5 años).

Se han propuesto dos opciones de asignación de rutas para el modelo avanzado.

Modelo avanzado. Opción A

En la primera de ellas, se asignan las rutas teniendo en cuenta el percentil que representa el valor de la OR agregada de cada fila, respecto de la muestra total generada. Los OR elevados implican mayor probabilidad de generar peores síntomas. Esto lo hemos operacionalizado asignando la ruta 1 cuando el percentil es menor o igual que la proporción de pacientes en R1. Asignaremos la ruta 2 si el percentil es mayor que la proporción de paciente en R1 y menor o igual que la suma de proporciones de R1 y R2. Será ruta 3 si el percentil es mayor que la suma de proporciones de R1 y R2, pero menor o igual que la suma de proporciones R1, R2 y R3. Será ruta 4 en todos los demás casos.

Las proporciones de pacientes en cada ruta (R1, R2, R3 o R4) las calculamos como una media ponderada, por el número de pacientes en cada rango de edad, de los valores de las columnas S,T, U, V de la Table 3.

Hay que tener en cuenta que las OR de la Table 5 no han sido ajustadas por la presencia de comorbilidades, que suelen ser más frecuentes conforme avanza la edad. Por ello, es probable que los datos generados con este modelo no sean del todo verosímiles, mostrando una presencia desmesurada de casos de edad avanzada en la ruta R4. Para resolver este problema planteamos la opción B.

Modelo avanzado. Opción B

En este caso, partimos de las mismas OR que en la opción A. Pero, se ha calculado el percentil que representa la OR de cada una de la fila solo comparando con otros casos dentro del mismo rango de edad (no de la muestra generada completa). La asignación de rutas se realizará también dentro de cada rango de edad en función de ese percentil y las proporciones de caso de las columnas S, T, U, V de la Table 3.

Paso 4

En el cuarto paso, se asigna a cada paciente la fecha de aparición de los síntomas severos y críticos (y, por lo tanto, el paciente necesitaría acudir a un hospital para ser ingresado en observación o en alguna de las otras unidades asistenciales). Para ello, se parte de una distribución de número de nuevos casos severos por día. Esta distribución puede generarse partiendo de datos reales. Por ejemplo, en los documentos que se adjuntan como *supplementary material* (https://zenodo.org/record/4699554), el número de casos que generan síntomas severos en nuestro data set respetan la proporción diaria de pacientes reales que han requerido ingreso hospitalario en la provincia de Valencia (España) entre el 1 de septiembre de 2020 y el 24 de marzo de 2021.

No obstante, se pueden usar otros conjuntos de datos de partida. Estos pueden ser reales, de otras áreas geográficas y/o ventanas temporales, o bien, pueden ser datos simulados partiendo de diferentes escenarios donde se diseñe una secuencia de una o más olas con diferente intensidad (pico de pacientes con síntomas severos en un determinado día) y amplitud (número de días desde el arranque de la ola hasta que se considera extinguida).

Para las filas que coincidan con cada categoría de *age range* y para cada *requirement path*, se ha generado el número de días que el paciente tendrá síntomas que requieran observación y oxígeno (DaysObs); el número de días en que requerirá NIMV (DaysNIMV); el número de días que requeriría IMV (DaysIMV); y, por último, el número de días en observación tras haber pasado por NIMV o IMV (DaysPstIMV). Para ello se ha usado distribuciones de Poisson con parámetro Lambda igual al promedio esperado de días con síntomas para ése rango de edad y ruta (Petermann-Rocha et al., 2020). Teniendo en cuenta que la duración de síntomas que requieren un tratamiento determinado es menor en las personas que fallecen que en las que se recuperan, se han parametrizado de forma diferente los pacientes de cada ruta. Esto ha supuesto una complejidad adicional, debido a la carencia de datos reales desglosados por rutas (Casas-Rojo et al., 2020; Guan et al., 2020; Huang et al., 2020; Rubio-Rivas et al., 2020; Wang et al., 2020; Xu et al., 2020). Hemos utilizado como parámetros los valores recogidos por un hospital grande español desglosados por rango de edad y resultado (*recovered* o *exitus*). En el momento de escribir este artículo no ha sido posible tener datos desglosados y diferentes para todos las ruta s (por lo que algunos parámetros se han repetido). Esto puede generar que los datos simulados tengan una ligera desviación respecto a la evolución de la enfermedad en los pacientes reales.

Por otra parte, se ha asumido una dsitribución poisson porque solo conocemos un parámetro (media). En el caso de disponer de más información para estimar *scale* and *shape*, podría ser interesante analizar en una investigación futura si una distribución Weibull podría ser más adecuada o no (Celeux et al., 2006; Epstein & Dexter, 2020; Mun, 2008).

Se ha considerado que la duración en días de cada síntoma es un *memoryless process*. De modo que, cada duración, es independiente de la duración de los otros estados de la enfermedad.

Guía para la reutilización de los datos

Cada conjunto de datos representa los datos demográficos, comorbilidades, fecha de ingreso y fechas de tránsito a otros estados (columnas) de un número de pacientes (filas).

Se pueden utilizar los datos generados para comprobar si un modelo, o un simulador, funciona al menos con datos prefabricados. Damos por hecho que, si el modelo no funciona con datos prefabricados, no funcionará con datos reales. El inverso no es necesariamente cierto, puede funcionar con datos prefabricados, pero fracasar con datos reales.

El algoritmo genera familias de conjuntos de datos (replicaciones) que comparten los mismos parámetros de generación, pero cambia la semilla de aleatorización (*seed*) en cada replicación.

Los parámetros a establecer para ejecutar el algoritmo son:

- El número de replicaciones a generar se puede establecer en el valor deseado (N)
- El número de pacientes a generar (filas del *data set*)
- *Age range*. Vector columna con el valor de edad que representa a cada intervalo de edad, normalmente el valor medio del intervalo
- Porcentaje de pacientes que se desea en cada *age range*. Vector columna con mismo número de elementos que rangos de edad contemplados. La suma de sus valores es igual a 100%
- Incidencia de la comorbilidad en la muestra. Porcentaje de pacientes que tendrán presente cada uno de los factores de riesgo contemplados
- Odds ratio de fallecer por Covid-19 que corresponde a la presencia de cada factor de riesgo respecto a la categoría de referencia (mujer, rango de edad 5 años, no embarazo y sin presencia de comorbilidades)
- Distribución de pacientes por rutas de requerimientos. Un vector columna por cada ruta considerado. Cada vector tiene tantos elementos como rangos de edad contemplados. La suma de los elementos de la misma fila será 100%
- Porcentaje de casos por día respecto al total de casos generados. Vector columna con tantos elementos como días contenga la ventana temporal a generar. Se pueden introducir números absolutos de casos por día (ya que el generador los convierte en porcentajes), o los porcentajes directamente. Se puede establecer una entrada constante cada día o los datos ajustados a la curva de entrada que se desee (variando picos, amplitud y olas)
- Los valores promedio de duración de cada síntoma, para cada rango de edad para cada ruta. Un vector columna por cada combinación de rutas y requerimientos asistenciales. En el ejemplo que mostramos, con 4 rutas (R1 a R4) y 4 requerimientos (observación, NIMV, IMV, Observación *post mechanical ventilation*), se precisan 16 vectores. Cada vector tiene tantos elementos como rangos de edad contemplados. Las ruta s donde no se produzcan determinados síntomas tendrán una duración de cero en todos los elementos de su vector. Si no se dispone de valores desagregados por edad y/o ruta se puede imputar el mismo valor promedio de duración en todos los elementos.

Se puede usar un conjunto de datos (o una parte del mismo) para calibrar los modelos. Posteriormente, se puede validar el modelo calibrado usando la otra parte del conjunto de datos (u otros conjuntos de datos de la misma o diferentes familias). De este modo, se puede comprobar si el modelo se ajusta adecuadamente a un conjunto de datos nuevos, no usados para refinar o calibrar el modelo. Si más delante se dispone de datos reales se podrá comprobar el rendimiento del modelo en un entorno real.

Por ejemplo, se puede genera una familia con N=30 conjuntos de datos, cada uno de ellos con 1500 filas. Se pude elaborar un modelo aprovechando las 1000 primeras filas de un data set y luego comprobar con las 500 filas restantes. O se puede elaborar el modelo con el data set completo y comprobarlo con los 29 data set de la familia o con data sets provenientes de otras familias.

Se pueden crear tantas familias de data sets como se considere, de modo que haya variaciones en los tiempos de duración de requerimientos (probando escenarios donde son totalmente aleatorios, o siguen algún tipo de distribución estadística o están en función de comorbilidades u otros parámetros del *data set*).

Los datos generados están basados en rutas de necesidades de los pacientes. En función de los recursos disponibles u otras políticas de salud, los pacientes pueden recibir el recurso que necesitan o no y eso

tiene unas claras consecuencias en el resultado final. Es decir, si un paciente de 85 años presenta *acute respiratory distress síndrome* (ARDS), precisaría de ventilación mecánica para sobrevivir y eso es lo que identificarán nuestro generador de datos. Otra cosa es que a ese paciente se le dispense ese tratamiento, o se decida que no compensa y se le pase a administrar cuidados paliativos. Por este motivo, los datos generados por nuestro algoritmo no tienen por qué cuadrar con los datos reales de pacientes ingresados en ICU (puesto que no todos los pacientes que necesitan ventilación mecánica la reciben).

Descripción del conjunto de datos

El documento ZIP que se ofrece como *supplementary material* (Figure 7) contiene el código de R en un notebook (extensión .RMD) y su versión en HTML. Además, incluye los archivos ".xlsx" de dos replicaciones generadas con los mismos parámetros. Cada replicación consta de tres data sets, uno para los datos generados con el modelo básico, otro con el modelo avanzado opción A y el tercero con el modelo avanzado opción B. Por eso, en cada replicación se generan tres *data sets* (dataset1bas, dataset2advA y dataset3advB). Los *data set* de una misma replicación comparten parámetros y semillas de aleatorización, pero asignan las ruta s de manera diferente. El modelo 1 no tiene en cuenta los parámetros de comorbilidad, ni de género; y asigna la ruta sólo en función del rango de edad para que el número de casos se parezca a la proporción de pacientes de cada ruta estipulados en los parámetros. El modelo 2 se basa en el riesgo combinado de comorbilidad, género y edad y solo se basa en eso para asignar rutas (ignorando la especificación de los parámetros donde se indica cuántos pacientes de cada rango de edad se desean en cada ruta). El tercer modelo asigna las ruta s teniendo en cuenta las *odds ratio* de todos los factores de riesgo (comorbilidad, género y rango de edad), pero estratifica por edades antes de repartir las rutas en base al riesgo. De este modo el modelo 3 respeta también las especificaciones del parámetro de distribución de pacientes por ruta según rango de edad.

El modelo más adecuado dependerá de la riqueza y fiabilidad de los datos introducidos como parámetros. Si apenas se dispone de información sobre la incidencia y/o OR de los factores de riesgo, pero si se tiene trazada la incidencia por rango de edad, el modelo 1 daría los resultados más parecidos a la realidad. Si no se tiene información de incidencia por rango de edad, pero las OR de los factores de riesgo son fiables, el modelo 2 sería la versión más adecuada. Cuando se disponen de datos completos, seguramente el modelo 3 representa mejor la realidad.

Cada replicación (rep) está etiquetada por el número que antecede a la extensión del fichero. Cada modelo de generación está etiquetado con un número a continuación de la palabra "Dataset" (1 para básico, 2 para opción A y 3 para opción B). La versión del algoritmo está etiquetada con el número a continuación de la "V". En el momento de publicar este trabajo la versión más actual es la versión 3.

El archivo del *notebook* de R está estructurado en tres secciones: 1) la declaración de las librerías necesarias; 2) la definición de parámetros (Figure 8); 3) el algoritmo de generación de datos basados en los parámetros (Figure 9)

La Figure 10 muestra los encabezados de columnas y las primeras filas del primero de los *datasets* generados (replicación 1, basic model). La descripción de cada una de las variables de la columna se

presenta en la Table 6. Cada *data set* se genera con las filas ordenadas por fecha de aparición de síntomas severos.

Para simplificar, se asumirá que todos los pacientes presentan todos los requerimientos (observación, VMNI, IMV, puesto de observación, ventilación mecánica y alta), pero la duración de los requerimientos no necesarios es cero (es decir, el día que muestra un requerimiento es igual al día del siguiente requerimiento). De esta forma, cuando sea necesario hacer cálculos con columnas de fecha, todas las filas tienen un día en cada columna. Incluso si el paciente no presenta necesidades IMV o post IMVneeds (en esos casos la fecha de alta es la misma que el día de IMVneeds o post IMVneeds. Por lo tanto, la duración de la estnacia en esos procesos es cero). Lo mismo para los pacientes con *exitus* en los que día de post IMVneeds es igual al día del alta

Futuras líneas de investigación usando este *data set*

Este trabajo abre líneas de investigación futura tanto teóricas como practicas. Desde el punto de vista teórico, seria interesante el desarrollo de herramientas de aprendizaje automático que, analizando muestras de datos especificas en hospitales reales, sean capaces de identificar los parámetros necesarios para el prototipado automático de generadores adaptados a cada hospital. En cuanto a las líneas de investigación aplicadas, es evidente que el formalismo propuesto para la generación de pacientes verosímiles no se limita a pacientes afectados por la infección del SARS-CoV-2. La generación de pacientes heterogéneos puede servir a representar las necesidades de una población especifica y servir de base para el estudio del comportamiento de sistemas de prestación de servicios de salud complejos.

Estos *data sets* pueden ser útiles para etapas previas de modelización de datos y comprobación de la viabilidad de modelos de eventos discretos, cuando los datos reales aún no están disponibles (por ejemplo, a la espera de aceptaciones de comités de ética o pendientes de recogida y procesamiento por parte de los centros de salud).

Los datos reales proporcionados por hospitales suelen ser incompletos o con inconsistencias. Por ello no son útiles para la validación de modelos de optimización o de simulación. El marco de trabajo presentado puede servir de punto de partida para los investigadores del área para extender o replicar el modelo.

En la medida que evolucione el sistema de información y traza de covid-19, se podría validar la verosimilitud de estos data sets y analizar posibles predictores par las ruta s no contemplados hasta la fecha.

Para las personas que ocupan cargos de gestión en hospitales, una ventaja que les proporciona esta investigación es que no tienen que partir desde cero. Es previsible que la aparición de variantes covid-19, el efecto de la vacunación y las nuevas tecnologías asistidas de respiración cambiarán los porcentajes de pacientes que circulen por cada trayectoria. Estos cambios se pueden modelizar fácilmente ajustando los parámetros y disponer rápidamente de un conjunto de datos que represente la nueva situación. Además, nuestra propuesta permite a las personas gestoras identificar qué información recolectar, almacenar y procesas en sus sistemas de información para dar un mejor soporte a las decisiones de gestión de camas covid-19.

Nuestros *data sets* también pueden servir para estimar las necesidades de los pacientes y diferenciarlo de los datos que reflejan las políticas de gestión de camas de hospital. De este modo, se podrían usar para analizar escenarios y valorar los efectos de determinadas políticas alternativas. O para analizar la sensibilidad de las decisiones de gestión de camas ante desviaciones de los parámetros de partida.

Por otra parte, durante la fase de parametrización del generador hay que alimentarlo con datos que, probablemente, no están disponibles y son susceptibles de ser objeto de investigación futura. Por ejemplo, el efecto las comorbilidades ajustadas por género y edad, o la incidencia de comorbilidades simultáneas.

También sería interesante comprobar hasta qué punto y en qué circunstancias, cada uno de los modelos de generación de datos proporciona unos resultados más verosímiles.

Del mismo modo, se podría ampliar el modelo incorporando más trayectorias o más etapas del proceso (por ejemplo, desde infecciones leves o asintomáticas y los requerimientos de atención domiciliaria).

Otra línea de investigación futura podría ocuparse de identificar valores de los parámetros para alimentar los modelos. Por ejemplo, estudiar la incidencia de las comorbilidades en determinadas poblaciones o la incidencia de varias comorbilidades simultáneamente en la misma persona. Otra mejora interesante sería la estimación de las OR respecto a la categoría de referencia, de necesitar NIMV, IMV o de fallecer ajustados por género, edad y otras comorbilidades. Es muy probable que las OR no sean independientes entre sí. Es decir, que el riesgo de sufrir determinadas comorbilidades está asociado a la edad o género. Los datos disponibles hasta la fecha no suelen estar ajustados a otras variables de posible confusión.

Por otra parte, este *data set* puede servir para comprobar el efecto de diversas decisiones de gestión hospitalaria que contribuyen a mejorar la experiencia del paciente o del personal sanitario. Por ejemplo, contar con mejores estimaciones de la ocupación de camas puede ayudar a mejorar la eficiencia de las decisiones en servicios auxiliares como lavandería, almacenes o cocinas; mejorar la gestión de almacenes o compras de suministros o equipos de protección sanitarios; o puede mejorar la eficiencia de la asignación de los turnos de personal sanitario.

Por último, se podría adaptar este generador a otras enfermedades diferente de la covid-19 pero que generan un consumo elevado de recursos asistenciales de los sistemas de salud, y donde sus trayectorias están bien estudiadas y permiten obtener fácilmente sus parámetros.

References

Alexander, G. L. (2007). The nurse-patient trajectory framework. *Medinfo. MEDINFO*, *12*(Pt 2), 910–914.

Belciug, S., Bejinariu, S. I., & Costin, H. (2020). An artificial immune system approach for a multi-compartment queuing model for improving medical resources and inpatient bed occupancy in pandemics. *Advances in Electrical and Computer Engineering*, *20*(3). https://doi.org/10.4316/AECE.2020.03003

BOE-A-2013-12887. (2013). *Ley 19/2013, de 9 de diciembre, de transparencia acceso a la información pública y buen gobierno.* 1–32.

Brochard, L. (2003). Mechanical ventilation: Invasive versus noninvasive. *European Respiratory Journal, Supplement*, *22*(47), 31s-37s. https://doi.org/10.1183/09031936.03.00050403

Buckley, D., & Gillham, M. (2007). Invasive Respiratory Support. In *Cardiothoracic Critical Care* (pp. 419–436). Elsevier Inc. https://doi.org/10.1016/B978-075067572-7.50032-1

Casas-Rojo, J. M., Antón-Santos, J. M., Millán-Núñez-Cortés, J., Lumbreras-Bermejo, C., Ramos-Rincón, J. M., Roy-Vallejo, E., Artero-Mora, A., Arnalich-Fernández, F., García-Bruñén, J. M., Vargas-Núñez, J. A., Freire-Castro, S. J., Manzano-Espinosa, L., Perales-Fraile, I., Crestelo-Viéitez, A., Puchades-Gimeno, F., Rodilla-Sala, E., Solís-Marquínez, M. N., Bonet-Tur, D., Fidalgo-Moreno, M. P., … Gómez-Huelgas, R. (2020). Clinical characteristics of patients hospitalized with COVID-19 in Spain: Results from the SEMI-COVID-19 Registry. *Revista Clinica Espanola*, *220*(8), 480–494. https://doi.org/10.1016/j.rce.2020.07.003

Castelnuovo, F., Marchese, V., Cristini, G., Crosato, V., Pennati, F., Renisi, G., Izzo, I., Paraninfo, G., Van Hauwermeiren, E., & Castelli, F. (2020). Discharge ward during the sars-cov-2 pandemic: An effective way to increase patient turnover when human resources are scarce. *Infezioni in Medicina*, *28*(4), 539–544.

Celeux, G., Lavergne, C., Vernaz, Y., Celeux, G., Lavergne, C., Vernaz, Y., Material, A., & Censored, D. (2006). *Assessing Material Aging from Doubly Censored Data : Weibull Distribution vs . Poisson Process To cite this version : HAL Id : inria-00072799 Assessing material aging from doubly censored data : Weibull distribution vs . Poisson process apport*. [Research Report] RR-3857, INRIA. 2000. inria-00072799.

Claudio, D., Cosgriff, V., Nino, V., & Valladares, L. (2021). An Agile Standardized Work Procedure for Cleaning the Operating Room. *Journal of Industrial Engineering and Management*, *14*, in press. https://doi.org/https://doi.org/jiem.3440

CNE -Centro Nacional de Epidemiología. (2020). *Información científico-técnica, enfermedad por coronavirus, COVID-19 (actualizado 20201112)*.

Comtois, D. (2021). *summarytools: Tools to Quickly and Neatly Summarize Data*.

Corbin, J. M., & Strauss, A. L. (1988). *Unending Work and Care: Managing Chronic Illness at Home*. Jossey-Bass Inc.

Daniel, P., Mecklenburg, M., Massiah, C., Joseph, M. A., Wilson, C., Parmar, P., Rosengarten, S., Maini, R., Kim, J., Oomen, A., & Zehtabchi, S. (2021). Non-invasive positive pressure ventilation versus endotracheal intubation in treatment of COVID-19 patients requiring ventilatory support. *American Journal of Emergency Medicine*, *43*, 103–108. https://doi.org/10.1016/j.ajem.2021.01.068

Dominguez-Lara, S. A. (2018). Odds-ratios and their interpretation as effect size in research. In *Educacion Medica* (Vol. 19, Issue 1, pp. 65–66). Fundacion Educacion Medica. https://doi.org/10.1016/j.edumed.2017.01.008

ECDP. (2020). Guidance for discharge and ending isolation in the context of widespread community transmission of COVID-19-first update Scope of this document. In *European Centre for Disease Prevention* (Issue April, pp. 1–8).

Epstein, R. H., & Dexter, F. (2020). A Predictive Model for Patient Census and Ventilator Requirements at Individual Hospitals During the Coronavirus Disease 2019 (COVID-19) Pandemic: A Preliminary Technical Report. *Cureus*. https://doi.org/10.7759/cureus.8501

European center for disease prevention and control. (2020). *Coronavirus disease 2019 (COVID-19) pandemic: increased transmission in the EU/EEA and the UK – seventh update*. *2019*(March).

Fowler, R., Hatchette, T., Salvadori, M., Baclic, O., Volling, C., Murthy, S., Emeriaud, G., Money, D., Brooks, J., Decou, M., & Ofner, M. (2020). Clinical management of patients with COVID-19:

Second interim guidance. *Canadian Critical Care Society and Association of Medical Microbiology and Infectious Disease (AMMI) Canada*, 1–67.

Garcia-Sabater, J. P., & Maheut, J. (2021). Introducción al Modelado Matematico, Nota Técnica. *RiuNet. Repositorio Institucional UPV*. https://doi.org/http://hdl.handle.net/10251/158555

Garcia-Sabater, J. P., Maheut, J., Ruiz, A., & Garcia-Sabater, J. J. (2020). A framework for capacity and operations planning in services organizations employing workers with intellectual disabilities. *Sustainability (Switzerland)*, *12*(22), 1–17. https://doi.org/10.3390/su12229713

Generalitat Valenciana. Conselleria de Sanitat Universal i Salut Pública. (2019). *Memoria de gestión conselleria de sanitat universal i salut pública 2019*. 14493–14496.

Generalitat Valenciana. (2018). *Memoria de Gestión de la Conselleria de Sanitat Universal i Salut Pública*.

Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei, C., Hui, D. S. C., Du, B., Li, L., Zeng, G., Yuen, K.-Y., Chen, R., Tang, C., Wang, T., Chen, P., Xiang, J., … Zhong, N. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China. *New England Journal of Medicine*, *382*(18), 1708–1720. https://doi.org/10.1056/NEJMoa2002032

Gunal, M. M. (2012). A guide for building hospital simulation models. *Health Systems*, *1*(1), 17–25. https://doi.org/10.1057/hs.2012.8

Hair, J. F., Black, W. C., Babin, B., & Anderson, R. E. (2009). *Multivariate data analysis (7th edition)*. Prentice Hall.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., … Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, *395*(10223), 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

Lagarda-Leyva, E. A., & Ruiz, A. (2019). A Systems Thinking Model to Support Long-Term Bearability of the Healthcare System: The Case of the Province of Quebec. *Sustainability*, *11*(24), 7028. https://doi.org/10.3390/su11247028

Manninen, K. (2020). *Typical progress of covid-19*.

Marin-Garcia, J. A. (2015). Publishing in two phases for focused research by means of "research collaborations." *WPOM-Working Papers on Operations Management*, *6*(2), 76. https://doi.org/10.4995/wpom.v6i2.4459

Marin-Garcia, J. A., Bonavia, T., & Losilla, J.-M. (2020). Changes in the Association between European Workers' Employment Conditions and Employee Well-Being in 2005, 2010 and 2015. *International Journal of Environmental Research and Public Health*, *17*(3), 1048. https://doi.org/10.3390/ijerph17031048

Marin-Garcia, J. A., Garcia-Sabater, J. P., Ruiz, A., Maheut, J., & Garcia-Sabater, J. J. (2020). Operations Management at the service of health care management: Example of a proposal for action research to plan and schedule health resources in scenarios derived from the COVID-19 outbreak. *Journal of Industrial Engineering and Management*, *13*(2), 213. https://doi.org/10.3926/jiem.3190

Marin-Garcia, J. A., Vidal-Carreras, P. I., Garcia Sabater, J. J., & Escribano-Martinez, J. (2019). Protocol: Value Stream Maping in Healthcare. A systematic literature review. *WPOM-Working Papers on Operations Management*, *10*(2), 36. https://doi.org/10.4995/wpom.v10i2.12297

Ministerio De Sanidad, Servicios Sociales e Igualdad. (2017). *Hábitos de Vida Informe Anual del Sistema Nacional de salud 2016* (INFORMES,). MINISTERIO DE SANIDAD, SERVICIOS SOCIALES E IGUALDAD.

Mun, J. (2008). Appendix C. Understanding and Choosing the Right Probability Distributions. *Advanced Analytical Models: Over 800 Models and 300 Applications from the Basel II Accord to Wall Street and Beyond*, 899–917. https://doi.org/10.1002/9781119197096.app03

Nino, V., Gomez, K., Martinez, K., & Claudio, D. (2021). Improving the registration process in a healthcare facility with lean principles. *Journal of Industrial Engineering and Management*, *14*, in press. https://doi.org/https://doi.org/jiem.3432

Olivieri, A., Palù, G., & Sebastiani, G. (2021). COVID-19 cumulative incidence, intensive care, and mortality in Italian regions compared to selected European countries. *International Journal of Infectious Diseases*, *102*. https://doi.org/10.1016/j.ijid.2020.10.070

Parlamento Europeo y del Consejo de la Unión Europea. (2019). *Directiva (UE) 2019/1024 DEL PARLAMENTO EUROPEO Y DEL CONSEJO de la Unión Europea de 20 de junio de 2019 relativa a los datos abiertos y la reutilización de la información del sector público (versión refundida)*. 172/56-172/78.

Petermann-Rocha, F., Hanlon, P., Gray, S. R., Welsh, P., Gill, J. M. R., Foster, H., Katikireddi, S. V., Lyall, D., Mackay, D. F., O'Donnell, C. A., Sattar, N., Nicholl, B. I., Pell, J. P., Jani, B. D., Ho, F. K., Mair, F. S., & Celis-Morales, C. (2020). Comparison of two different frailty measurements and risk of hospitalisation or death from COVID-19: findings from UK Biobank. *BMC Medicine*, *18*(1). https://doi.org/10.1186/s12916-020-01822-4

Pinaire, J., Azé, J., Bringay, S., & Landais, P. (2017). Patient healthcare trajectory. An essential monitoring tool: a systematic review. *Health Information Science and Systems*, *5*(1), 1–18. https://doi.org/10.1007/s13755-017-0020-2

Plaza, J. (2021). Informe Científico-Divulgativo: Un Año De Coronavirus Sars-Cov-2. *Miniserio de Ciencia e Innovación*.

Popat, B., & Jones, A. T. (2012). Invasive and non-invasive mechanical ventilation. In *Medicine (United Kingdom)* (Vol. 40, Issue 6, pp. 298–304). Elsevier Ltd. https://doi.org/10.1016/j.mpmed.2012.03.010

Posso, M., Comas, M., Román, M., Domingo, L., Louro, J., González, C., Sala, M., Anglès, A., Cirera, I., Cots, F., Frías, V.-M., Gea, J., Güerri-Fernández, R., Masclans, J. R., Noguès, X., Vázquez, O., Villar-García, J., Horcajada, J. P., Pascual, J., & Castells, X. (2020). Comorbidities and Mortality in Patients With COVID-19 Aged 60 Years and Older in a University Hospital in Spain. *Archivos de Bronconeumología*, *56*(11), 756–758. https://doi.org/10.1016/j.arbres.2020.06.012

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*.

Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research*.

Roa-Martínez, S. M., Vidotti, S. A. B., & Santana, R. C. (2017). Estructura propuesta del artículo de datos como publicación científica. *Revista Espanola de Documentacion Cientifica*, *40*(1), 1–12. https://doi.org/10.3989/redc.2017.1.1375

Romeo Casabona, C. M., & Urruela Mora, A. (2020). *Informe Del Ministerio De Sanidad Sobre Los Aspectos Éticos En Situaciones De Pandemia: El Sars-Cov-2*. 12.

RStudio Team. (2020). *RStudio: Integrated Development for R*. RStudio, PBC.

Rubio-Rivas, M., Corbella, X., Mora-Luján, J. M., Loureiro-Amigo, J., López Sampalo, A., Yera Bergua, C., Esteve Atiénzar, P. J., Díez García, L. F., Gonzalez Ferrer, R., Plaza Canteli, S., Pérez Piñeiro, A., Cortés Rodríguez, B., Jorquer Vidal, L., Pérez Catalán, I., León Téllez, M., Martín Oterino, J. Á., Martín González, M. C., Serrano Carrillo de Albornoz, J. L., García Sardon, E., … Gómez-Huelgas, R. (2020). Predicting Clinical Outcome with Phenotypic Clusters in COVID-19

Pneumonia: An Analysis of 12,066 Hospitalized Patients from the Spanish Registry SEMI-COVID-19. *Journal of Clinical Medicine*, *9*(11), 3488. https://doi.org/10.3390/jcm9113488

Ruckdeschel, P., Kohl, M., Stabla, T., & Camphausen, F. (2006). S4 Classes for Distributions. *R News*, *6*(2), 2–6.

Ruza, F. (2008). *Cuidados Intensivos Pediatricos*. *6*(6), 336.

Schauberger, P., & Walker, A. (2020). *openxlsx: Read, Write and Edit xlsx Files*.

Stang, A., Stang, M., & Jöckel, K. H. (2020). Estimated use of intensive care beds due to COVID-19 in Germany over time. *Deutsches Arzteblatt International*, *117*(19). https://doi.org/10.3238/arztebl.2020.0329

Unroe, M., Kahn, J. M., Carson, S. S., Govert, J. A., Martinu, T., Sathy, S. J., Clay, A. S., Chia, J., Gray, A., Tulsky, J. A., & Cox, C. E. (2010). One-year trajectories of care and resource utilization for recipients of prolonged mechanical ventilation: A cohort study. *Annals of Internal Medicine*, *153*(3), 167–175. https://doi.org/10.7326/0003-4819-153-3-201008030-00007

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). Springer.

Wang, Y., Wang, Y., Chen, Y., & Qin, Q. (2020). Unique epidemiological and clinical features of the emerging 2019 novel coronavirus pneumonia (COVID-19) implicate special control measures. In *Journal of Medical Virology* (Vol. 92, Issue 6, pp. 568–576). John Wiley and Sons Inc. https://doi.org/10.1002/jmv.25748

Wickham, H. (2007). Reshaping Data with the {reshape} Package. *Journal of Statistical Software*, *21*(12), 1–20.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, *40*(1), 1–29.

Wiersema, U. F. (2007). Noninvasive Respiratory Support. In *Cardiothoracic Critical Care* (pp. 410–418). Elsevier Inc. https://doi.org/10.1016/B978-075067572-7.50031-X

Winck, J. C., & Scala, R. (2021). Non-invasive respiratory support paths in hospitalized patients with COVID-19: proposal of an algorithm. *Pulmonology*. https://doi.org/10.1016/j.pulmoe.2020.12.005

Wong, G. N., Weiner, Z. J., Tkachenko, A. V., Elbanna, A., Maslov, S., & Goldenfeld, N. (2020). Modeling COVID-19 dynamics in Illinois under non-pharmaceutical interventions. In *medRxiv*. https://doi.org/10.1101/2020.06.03.20120691

Wu, H., Godfrey, A. J. R., Govindaraju, K., & Pirikahu, S. (2020). *ExtDist: Extending the Range of Functions for Probability Distributions*.

Xia, W., & Sun, J. (2013). Simulation guided value stream mapping and lean improvement: A case study of a tubular machining facility. *Journal of Industrial Engineering and Management*, *6*(2), 456–476. https://doi.org/10.3926/jiem.532

Xu, X. W., Wu, X. X., Jiang, X. G., Xu, K. J., Ying, L. J., Ma, C. L., Li, S. B., Wang, H. Y., Zhang, S., Gao, H. N., Sheng, J. F., Cai, H. L., Qiu, Y. Q., & Li, L. J. (2020). Clinical findings in a group of patients infected with the 2019 novel coronavirus (SARS-Cov-2) outside of Wuhan, China: Retrospective case series. *The BMJ*, *368*. https://doi.org/10.1136/bmj.m606

Zheng, Z., Peng, F., Xu, B., Zhao, J., Liu, H., Peng, J., Li, Q., Jiang, C., Zhou, Y., Liu, S., Ye, C., Zhang, P., Xing, Y., Guo, H., & Tang, W. (2020). Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. In *Journal of Infection* (Vol. 81, Issue 2, pp. e16–e25). W.B. Saunders Ltd. https://doi.org/10.1016/j.jinf.2020.04.021