Understanding COPD patients in the hospital system via administrative data

Henry Wilde, Vincent Knight, Jonathan Gillard

Abstract

This work presents an analysis of how patients with chronic obstructive pulmonary disorder (COPD) interact with the hospital system in South Wales.

1 Introduction

This introduction will briefly summarise the literature review for studying a patient corpus via clustering. Following this, a condensed data analysis is presented highlighting the main conclusions of the clustering and the overall benefits compared with traditional condition-treatment segmentation.

2 Constructing the queuing model

Owing to a lack of available data on the system and its patients, the options for the queuing model used are limited compared to those employed in some modern works. However, there is a precedent for simplifying healthcare systems to a single node with, for example, parallel servers that emulate resource availability. [3] and [5] provide good examples of how this approach, when paired with discrete event simulation, can expose the resource needs of a system beyond deterministic queuing theory models. In particular, [5] shows how a single node, multiple server queue can be used to accurately predict bed capacity and length of stay distributions in a critical care unit using administrative data.

To follow in the suit of recent literature, a single node using a $M|M|c$ queue is employed to model a hypothetical ward of patients presenting COPD. In addition to this, the grouping found in Section **??** provides a set of patient classes in the queue. Under this model, the following assumptions are made:

1. Inter-arrival and service times of patients are each exponentially distributed with some mean. This is in spite of the system time distributions shown in Figure **??** in order to simplify the model parameterisation.

2. There are c servers available to arriving patients at the node representing the overall resource availability including bed capacity and medical staff.

3. There is no queue or system capacity. In [5], a queue capacity of zero is set under the assumption that any surplus arrivals would be sent to another suitable ward or unit. As this hypothetical ward represents COPD patients potentially throughout a hospital, this assumption is not held.

4. Without the availability of expert clinical knowledge, a first-in first-out service policy is employed in lieu of some patient priority framework.

Each group of patients has its own arrival distribution. The parameter of this distribution is taken to be the reciprocal of the mean inter-arrival times for that group.

Like arrivals, each group of patients has its own service time distribution. Without full details of the process order or idle periods during a spell, some assumption must be made about the true 'service' time of a patient in hospital. It is assumed here that the mean service time of a group of patients may be approximated via their mean length of stay, i.e. the mean time spent in the system. For simplicity, this work considers the mean service time, $\frac{1}{\mu}$, to be directly proportional to the mean total system time, $\frac{1}{\phi}$, such that:

$$\mu = p\phi \tag{1}$$

where $p \in (0, 1]$ is some parameter to be determined for each group, denoted by p_i for group i.

One of the few ground truths available in the provided data is the distribution of the total length of stay. Given that the length of stay and resource availability are connected, the approach here will be to simulate the length of stay distribution for a range of values p_i and c in order to find the parameters that best match the observed data.

The statistical comparison of two or more distributions can be done in a number of ways. Such methods include the Kolmogorov-Smirnov test, a variety of discrepancy approaches such as summed mean-squared error, and f-divergences. A popular choice amongst the latter group (which may be considered distance-like) is the Kullback-Leibler divergence which measures relative information entropy from one probability distribution to another [1]. The key issue with many of these methods is that they lack interpretability which is paramount when conveying information to stakeholders. Interpretability not just from explaining how something works but how its results may be explained also.

As such, a reasonable candidate is the (first) Wasserstein metric, also known as the 'earth mover' or 'digger' distance [4]. The Wasserstein metric satisfies the conditions of a formal mathematical metric (like the typical Euclidean distance), and its values take the units of the distributions under comparison (in this case: days). Both of these characteristics can aid understanding and explanation. In simple terms, the distance measures the approximate 'minimal work' required to move between two probability distributions where 'work' can be loosely defined as the product of how much of the distribution's mass is to be moved and the distance it must be moved by. More formally, the Wasserstein distance between two probability distributions U and V is defined as:

$$W(U, V) = \int_0^1 \left| F^{-1}(t) - G^{-1}(t) \right| dt \tag{2}$$

where F and G are the cumulative density functions of U and V respectively. A proof of (2) is presented in [2]. The parameter set with the smallest maximum distance between any cluster's simulated system time distribution and the overall observed length of stay distribution is then taken to be the most appropriate.

To be specific, let T denote the system time distribution of all of the observed data and let $T_{i,c,p}$ denote the system time distribution for cluster i obtained from a simulation with c servers and $p := (p_0, p_1, p_2, p_3)$.

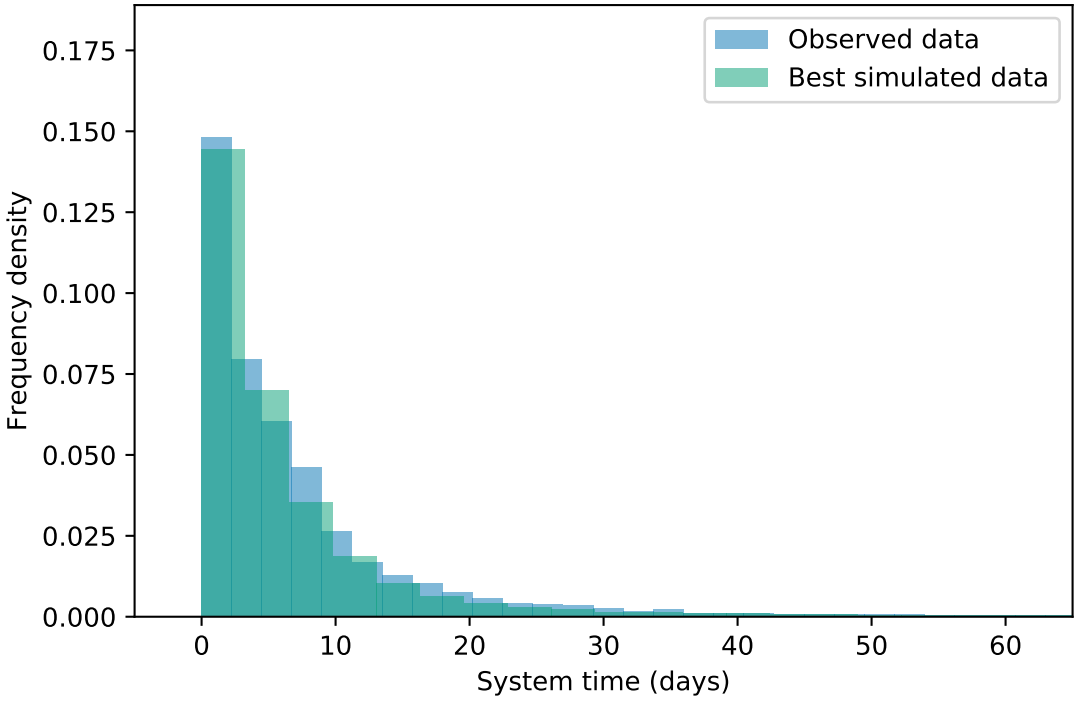

Figure 1: A histogram of the simulated and observed length of stay data for the best parameter set.

Then the optimal parameter set (c^*, p^*) is given by:

$$(c^*, p^*) = \arg\min_{c,p} \left\{ \max_i \{W(T_{i,c,p}, T)\} \right\} \tag{3}$$

The parameter sweep included values of each p_i from 0.5 to 1 with a granularity of 5×10^{-2} and values of c from 40 to 60 at steps of 5. These choices were informed by the assumptions of the model and formative analysis to reduce the parameter space given the computational resources required to conduct the simulations. Each parameter set was repeated 50 times with each simulation running for four years of virtual time. The warm-up and cool-down periods were taken to be approximately one year each leaving two years of simulated data from each repetition.

The results of this parameter sweep can be summarised in Figures 1 and 2. Each figure shows a comparison of the observed lengths of stay across all groups and the newly simulated data with the best and worst parameter sets respectively. It can be seen that, in the best case, a very close fit has been found. Meanwhile, Figure 2 highlights the importance of good parameter estimation under this model since the likelihood of short-stay patient arrivals has been inflated disproportionately against the tail of the distribution. Table 1 reinforces these results numerically, showing a clear fit by the best parameters across the board.

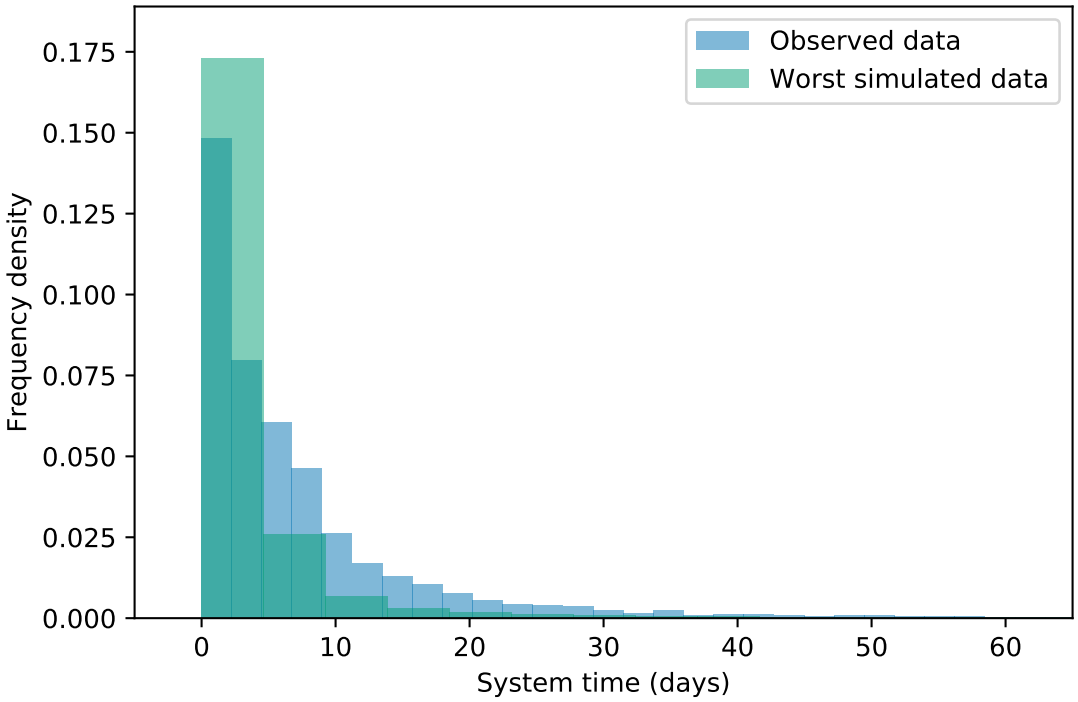

Figure 2: A histogram of the simulated and observed length of stay data for the worst parameter set.

	Model parameter and result						LOS statistic						
	p_0	p_1	p_2	p_3	c	Max. distance	Mean	Std.	Min.	25%	Med.	75%	Max.
Observed	NaN	NaN	NaN	NaN	NaN	0.00	7.70	11.86	-0.02	1.49	4.20	8.93	224.93
Best simulated	0.95	1.0	1.0	0.5	40.0	1.28	7.00	12.09	0.00	1.44	3.57	7.65	326.46
Worst simulated	0.50	0.5	0.5	1.0	40.0	4.25	4.36	13.40	0.00	0.72	1.78	3.84	463.01

Table 1: A comparison of the observed data, and the best and worst simulated data based on the model parameters and summary statistics for length of stay (LOS).

3 Adjusting the queuing model

Body of the writing and plots come here. What can we see in the what-if scenarios? The main scenarios are:

- How would server utilisation (i.e. resource consumption) be affected by an increase in overall patient arrivals?
- How is the system affected by certain types of patients (e.g. short-stay, low-impact) arriving less frequently?
- What are the sensitivities of mean system times and server utilisation based on a change in c?

4 Conclusion

Summarise the findings and novelty of the paper: sensitivity analysis and queuing models are within reach despite a lack of data. The chosen modelling discipline for service times is very simplistic but can return good results (refer back to best-case parameter plot).

References

[1] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1): 79–86, 1951. doi: 10.1214/aoms/1177729694.

[2] A. Ramdas, N. G. Trillos, and M. Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017. doi: 10.3390/e19020047.

[3] K. Steins and S. Walther. A generic simulation model for planning critical care resource requirements. *Anaesthesia*, 68(11):1148–1155, 2013.

[4] L. N. Vaserstein. Markov processes over denumerable products of spaces describing large systems of automata. *Problemy Peredačhi Informatsii*, 5(3):64–72, 1969.

[5] J. Williams, S. Dumont, J. Parry-Jones, I. Komenda, J. Griffiths, and V. Knight. Mathematical modelling of patient flows to predict critical care capacity required following the merger of two district general hospitals into one. *Anaesthesia*, 70(1):32–40, 2015. doi: 10.1111/anae.12839.