Segmentation analysis and the recovery of queuing parameters via the Wasserstein distance: a study of administrative data for patients with chronic obstructive pulmonary disease

Henry Wilde[a], Vincent Knight[a], Jonathan Gillard[a] and Kendal Smith[b]

[a]School of Mathematics, Cardiff University, UK; [b]Cwm Taf Morgannwg University Health Board, UK

ARTICLE HISTORY
Compiled March 14, 2022

ABSTRACT

This work uses a data-driven approach to analyse how the resource requirements of patients with chronic obstructive pulmonary disease (COPD) may change, quantifying how those changes impact the hospital system with which the patients interact. This approach is composed of a novel combination of often distinct modes of analysis: segmentation, operational queuing theory, and the recovery of parameters from incomplete data. By combining these methods as presented here, this work demonstrates that potential limitations around the availability of fine-grained data can be overcome. Thus, finding useful operational results despite using only administrative data.

The paper begins by finding a useful clustering of the population from this granular data that feeds into a multi-class $M/M/c$ model, whose parameters are recovered from the data via parameterisation and the Wasserstein distance. This model is then used to conduct an informative analysis of the underlying queuing system and the needs of the population under study through several what-if scenarios.

The analyses used to form and study this model consider, in effect, all types of patient arrivals and how those types impact the system. With that, this study finds that there are no quick solutions to reduce the impact of COPD patients on the system, including adding capacity to the system. In this analysis, the only effective intervention to reduce the strain caused by those presenting with COPD is to enact external policies which directly improve the overall health of the COPD population before they arrive at the hospital.

KEYWORDS
OR in health services; machine learning; queueing

1. Introduction

Population health research is increasingly based on data-driven methods for patient-centred care — as opposed to those designed solely by clinical experts. This movement is borne from the advent of accessible software and a relative abundance of electronic data. However, many such methods rely heavily on detailed data about both the healthcare system and its population, which may limit research where sophisticated data pipelines are not yet in place.

The only healthcare datasets used in this work are administrative hospital records.

CONTACT Henry Wilde. Email: henrydavidwilde@gmail.com

These records offer little detail as to the exact nature of a patient's time in hospital other than a surface-level summary. Exploratory analysis of a population-wide administrative dataset, omitted from this manuscript, revealed the presence of high variation in almost all regards. This variability stifles the possibility of uncovering valuable insights about the whole population. However, some benefits are made apparent by considering a condition-specific population. This work utilises another administrative dataset of patients presenting COPD, and demonstrates how actionable insights can be identified by thoroughly extracting information from the dataset through the use of machine learning and operational research techniques.

This manuscript presents a method of overcoming this, using routinely gathered, administrative hospital data to build a clustering which feeds into a multi-class queuing model, allowing for better understanding of the healthcare population and the system with which they interact. COPD is a condition of particular interest to population health research, and to Cwm Taf Morgannwg University Health Board, as it is known to often present as a comorbidity in patients Houben-Wilke et al. (2019), increasing the complexity of treatments among those with the condition. Moreover, an internal report by NHS Wales found Cwm Taf Morgannwg University Health Board had the highest prevalence of the condition across all the Welsh health boards.

This work draws upon several overlapping sources within mathematical research, and this work contributes to the literature in three ways: to theoretical queuing research by the estimation of missing queuing parameters with the Wasserstein distance; to operational healthcare research through the weaving together of the combination of methods used in this work despite data constraints; and to public health research by adding to the growing body of mathematical and operational work around a condition that is vital to understand operationally, socially and medically.

The remainder of this manuscript is structured as follows:

- Section 1 provides a brief literature review, followed by an overview of the dataset and its clustering;
- Section 2 offers a concise introduction to queuing theory;
- Section 3 describes the queuing model and the estimation of its parameters;
- Section 4 presents several what-if scenarios with insight provided by the model parameterisation and the clustering;
- Section 5 summarises the manuscript and its findings.

1.1. Literature review

Given the subject matter of this work, the relevant literature spans much of operational research in healthcare, and the focus of this review is on the critical topics of segmentation analysis, queuing models applied to hospital systems, and the handling of missing or incomplete data for such queues.

1.1.1. Segmentation analysis

Segmentation analysis allows for the targeted analysis of otherwise heterogeneous datasets and encompasses several techniques from operational research, statistics and machine learning. One of the most desirable qualities of this kind of analysis is the ability to glean and communicate simplified summaries of patient needs to stakeholders within a healthcare system Vuik, Mayer, and Darzi (2016a); Yoon, Goh, Kwan, Thumboo, and Low (2020). For instance, clinical profiling often forms part of the broader analysis where each segment is summarised in a phrase or infographic Vuik, Mayer, and Darzi (2016b); Yan et al. (2019).

The review for this work identified three commonplace groups of patient characteristics used to segment a patient population: system utilisation metrics; clinical attributes; and the pathway. The last is not used to segment the patients directly, instead of grouping their movements through a healthcare system, typically via process mining. Arnolds and Gartner (2018) and Delias, Doumpos, Grigoroudis, Manolitzas, and Matsatsinis (2015) demonstrate how this technique can be used to improve the efficiency of a hospital system as opposed to tackling the more relevant issue of patient-centred care. The remaining characteristics can be segmented in a variety of ways, but recent works tend to favour unsupervised methods —– typically latent class analysis (LCA) or clustering Yan, Kwan, Tan, Thumboo, and Low (2018).

LCA is a statistical, model-based method used to identify groups (called latent classes) in data by relating its observations to some unobserved (latent), categorical attribute. This attribute has multiple possible categories, each corresponding to a latent class. The discovered relations enable the observations to be separated into latent classes according to their maximum likelihood class membership Hagenaars (2002); Lazarsfeld and Henry (1968). This method has proved useful in the study of comorbidity patterns as in Kuwornu, Lix, and Shooshtari (2014); Larsen, Pedersen, Friis, Glümer, and Lasgaard (2017) where combinations of demographic and clinical attributes are related to various subgroups of chronic diseases.

Similarly to LCA, clustering identifies groups (clusters) in data to produce labels for its instances. However, clustering includes a wide variety of methods where the common theme is to maximise homogeneity within, and heterogeneity between, each cluster Everitt, Landau, Leese, and Stahl (2011). The k-means paradigm is the most popular form of clustering in literature. The method iteratively partitions numerical data into $k \in \mathbb{N}$ distinct parts where k is fixed a priori. This method has proved popular as it is easily scalable, and its implementations are concise Olafsson, Li, and Wu (2008); Wu and Kumar (2009). In addition to k-means, hierarchical clustering methods can be useful if a suitable number of parts cannot be found initially Vuik et al. (2016b). However, supervised hierarchical segmentation methods such as classification and regression trees (as in Harper and Winslett (2006)) have been used where an existing, well-defined, label is of particular significance.

1.1.2. Queuing models

Since the seminal works by Erlang Erlang (1917, 1920) established the core concepts of queuing theory, the application of queues and queuing networks to real services has become abundant, including the healthcare service. By applying these models to healthcare settings, many aspects of the underlying system can be studied. A common area of study in healthcare settings is of service capacity. McClain (1976) is an early example of such work where acute bed capacity was determined using hospital occupancy data. Meanwhile, more modern works such as Palvannan and Teow (2012); Pinto, de Campos, Perpétuo, and Ribeiro (2014) consider more extensive sources of data to build their queuing models. Moreover, the output of a model is catered more towards being actionable — as is the prerogative of operational research. For instance, Pinto et al. (2014) devises new categorisations for both hospital beds and arrivals that are informed by the queuing model. A further example is Komashie, Mousavi, Clarkson, and Young (2015) where queuing models are used to measure and understand satisfaction among patients and staff.

In addition to these theoretic models, healthcare queuing research has expanded to include computer simulation models. The simulation of queues, or networks thereof, have the benefit of adeptly capturing the stochastic nuances of hospital systems over their theoretic counterparts. Example areas include the construction and simulation of Markov processes via process mining Arnolds and Gartner (2018); Rebuge and Ferreira (2012), and patient flow Bhattacharjee and Ray (2014). Regardless of the advantages of simulation models, a prerequisite is reliable software with which to construct those simulations. A common approach to building simulation models of queues is to use a graphical user interface such as Simul8. These tools have the benefits of being highly visual, making them attractive to organisations looking to implement queuing models without necessary technical expertise, including the NHS. Brailsford et al. (2013) discusses the issues around operational research and simulation being taken up in the NHS despite the availability of intuitive software packages like Simul8. However, they do not address a core principle of good simulation work: reproducibility. The ability to reliably reproduce a set of results is of great importance to scientific research but remains an issue in simulation research generally Fitzpatrick (2019). When considering issues with reproducibility in scientific computing (simulation included), the source of any concerns is often with the software used Ivie and Thain (2018). Using well-developed, open-source software can alleviate issues around reproducibility and reliability as how they are used involve less uncertainty and require more rigour than ‘drag-and-drop’ software. One example of such a piece of software is Ciw Palmer, Knight, Harper, and Hawa (2019). Ciw is a discrete event simulation library written in Python that is fully documented and tested. The simulations constructed and studied in Sections 3 and 4 utilise this library and aid the overall reproducibility of this work.

1.1.3. Handling incomplete queue data

As is discussed in other parts of this section, the data available in this work is not as detailed as in other comparative works. Without access to such data — but intending to gain insight from what is available — it is imperative to bridge the gap left by the incomplete data.

Moreover, it is often the case that in practical situations where suitable data is not (immediately) available, further inquiry in that line of research will stop. Queuing models in healthcare settings appear to be such a case; the line ends at incomplete queue data. Asanjarani, Nazarathy, and Pollett (2017) is a bibliographic work that collates articles on the estimation of queuing system characteristics — including their parameters. Despite its breadth of almost 300 publications from 1955, only two articles have been identified as being applied to healthcare: Mohammadi and Salehi-Rad (2012); Yom-Tov and Mandelbaum (2014). Both works are concerned with customers who can re-enter services during their time in the queuing system, which is mainly of value when considering the effect of unpredictable behaviour in intensive care units, for instance. Mohammadi and Salehi-Rad (2012) seeks to approximate service and re-service densities through a Bayesian approach and by filtering out those customers seeking to be serviced again. On the other hand, Yom-Tov and Mandelbaum (2014) considers an extension to the $M/M/c$ queue with direct re-entries. The devised model is then used to determine resource requirements in two healthcare settings.

Aside from healthcare-specific works, the approximation of queue parameters has formed a part of relevant modern queuing research. However, the scope is primarily focused on theoretic approximations rather than by simulation. Djabali, Rabta, and Aissani (2018); Goldenshluger (2016) are two such recent works that consider an un-

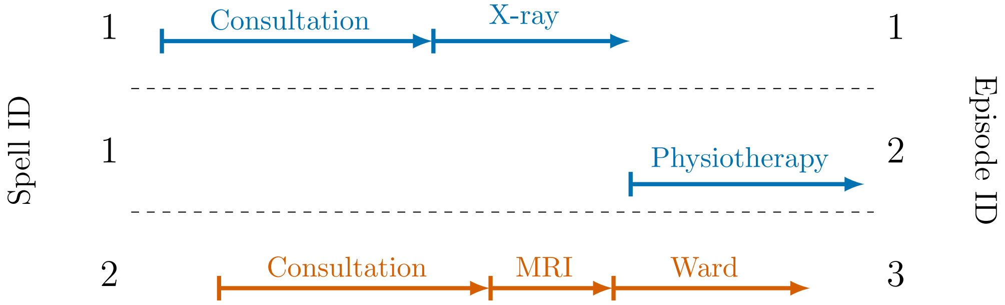


Figure 1. A Gantt chart of two patient spells across three episodes

derlying process to estimate a general service time distribution in single server and infinite server queues respectively.

1.2. Overview of the dataset and its clustering

Cwm Taf Morgannwg University Health Board provided the dataset used in this work. The dataset contains an administrative summary of 5,231 patients presenting COPD from February 2011 through to March 2019, covering 10,861 hospital spells. A patient (hospital) spell is defined as the continuous stay of a patient using a hospital bed on premises controlled by a healthcare provider, and is made up of one or more patient episodes NHS Data Model and Dictionary (n.d.-b). A patient episode is defined to be any continuous period of care provided by the same consultant NHS Data Model and Dictionary (n.d.-a). Figure 1 contains an illustrative example of the relationship between patient episodes and patient spells.

In the example, the first patient begins their first episode with a consultation and an X-ray. Following this, they are referred to a physiotherapist for specialist treatment; this is their second episode, and concludes their spell. The second patient spell consists of consultation, an MRI, and monitoring on a ward. These are all overseen by the same consultant, and so they form one episode. The analysis in Appendix **??** considers another dataset consisting of patient episodes. However, as is often the case with administrative datasets, the order of procedures within the same episode is not recorded.

The following attributes describe the spells included in the dataset studied in this work:

- Personal identifiers and information, i.e. patient and spell ID numbers, and identified gender;
- Admission/discharge dates and approximate times;
- Attributes summarising the clinical path of the spell including admission/discharge methods, and the number of episodes, consultants and wards in the spell;
- International Classification of Diseases (ICD) codes and primary Healthcare Resource Group (HRG) codes from each episode;
- Indicators for any COPD intervention. The value for any given instance in the dataset (i.e. a spell) is one of no intervention, pulmonary rehabilitation (PR), specialist nursing (SN), and both interventions;
- Charlson Comorbidity Index (CCI) contributions from several long term condi-

tions (LTCs) as well as indicators for some other conditions such as sepsis and obesity. CCI is useful in anticipating hospital utilisation as a measure for the burdens associated with comorbidity Simon-Tuval et al. (2011);
- Rank under the 2019 Welsh Index of Multiple Deprivation (WIMD), indicating relative deprivation of the postcode area the patient lives in which is known to be linked to COPD prevalence and severity Collins, Stratton, Kurukulaaratchy, and Elia (2018); Sexton and Bedford (2016); Steiner et al. (2017).

In addition to the above, the following attributes were engineered for each spell:

- Age and spell cost data were linked to approximately half of the spells in the dataset from the administrative dataset analysed in Appendix **??**;
- The presenting ICD codes were generalised to their categories according to NHS documentation and counts for each category were attached. This reduced the number of values from 1,926 codes to 21 categories;
- A measure of admission frequency was calculated by taking the number of COPD-related admissions in the last twelve months linked to the associated patient ID number.

Although there is a fair amount of information here, it is limited to COPD-related admissions. Therefore, rather than segmenting the patients themselves, the spells will be segmented.

The attributes included in the clustering encompass both utilisation metrics and clinical attributes relating to the spell. They comprise the summary clinical path attributes, the CCI contributions and condition indicators, the WIMD rank, length of stay (LOS), COPD intervention status, and the engineered attributes (not including age and costs due to lack of coverage between the two datasets).

With these attributes selected, a clustering algorithm must be chosen. Two critical specifications of the algorithm used are that it must handle mixed-type data, and that it should be interpretable by stakeholders. As such, the k-prototypes algorithm is a strong candidate. The k-prototypes algorithm was mentioned in Chapter **??** and is a mixed-type extension to the k-modes and k-means algorithms; in effect, the k-prototypes algorithm separates the given dataset into its numeric and categorical attributes before applying k-means and k-modes on the respective parts. The statement of the k-prototypes algorithm has been omitted since it is equivalent to that of k-modes (given in Algorithm **??**) with the exceptions that:

- The SelectClosest function uses the dissimilarity measure given in (2);
- The Update function is as given in Algorithm 1.

These parts are combined using a modified dissimilarity function, defined in (2). This function is a linear combination of the squared Euclidean distance and the dissimilarity function defined in (**??**) according to a weight, $\gamma \in \mathbb{R}$. The notation and terminology for clustering mixed-type data is much the same as in Chapter **??**. However, there are substantial differences: first, representative points are referred to as *prototypes*; and, second, an attribute space $\mathcal{A}$ of m mixed-type attributes can be written as the product of its numeric and categorical components:

$$\mathcal{A} = \prod_{j=1}^{p} A_j^{(n)} \times \prod_{j=p+1}^{m} A_j^{(c)} = \mathcal{A}^{(n)} \times \mathcal{A}^{(c)} \tag{1}$$

Here, $A^{(n)}$ and $A^{(c)}$ denote individual numeric and categorical attributes, respectively. Meanwhile, $\mathcal{A}^{(n)}$ and $\mathcal{A}^{(c)}$ denote the numeric and categorical components of the space. With this notation, the dissimilarity between two points, $X, Y \in \mathcal{A}$, is defined to be:

$$d(X,Y) = \sum_{j=1}^{p} (x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_j, y_j) \tag{2}$$

```
Algorithm 1: UPDATE (k-prototypes)
  Input: an attribute space $\mathcal{A} = \mathcal{A}^{(n)} \times \mathcal{A}^{(c)}$, a prototype to update $z^{(l)}$ and its
         cluster $Z_l$
  Output: an updated prototype
1 Find $z_n \in \mathcal{A}^{(n)}$, the mean numeric attribute vector in $Z_l$:
      $z_n := \left( \frac{1}{|Z_l|} \sum_{u \in Z_l} u_j : j = 1, \dots, p \right)$
2 Find $z_c \in \mathcal{A}^{(c)}$ that minimises $D\left(Z_l^{(c)}, z_c\right)$ where:
      $Z_l^{(c)} := \{(u_j : j = p+1, \dots, m) : u \in Z_l\}$
   i.e. find the modal categorical attribute vector in $Z_l$
3 Update the prototype to be the concatenation of these vectors:
      $z^{(l)} \leftarrow z_n \frown z_c$
```

In addition to this dissimilarity function, k-prototypes has a cost function that uses the same linear combination as its dissimilarity function to consider the numeric and categorical attributes. This function combines the categorical cost function (defined in (**??**)) and inertia (defined in (**??**)) according to the same value of γ. A proof that minimising this linear combination also minimises the intra-cluster k-prototypes dissimilarity is given in Huang (1997).

The choice of γ is of particular importance as it balances the contribution of each data type to the objective function. The seminal work by Huang Huang (1997) investigated the effect of various γ values when clustering with k-prototypes. This investigation determined that a sensible and robust value for γ is the average of the standard deviations for the numeric attributes. The analysis that informed the clustering in this work found that this value for γ provided a useful clustering; as such, no further modifications were made.

To determine the optimal number of clusters, k, the knee point detection algorithm used at the end of Chapter **??** was used with a range of potential values for k from two to ten. This range was chosen based on what may be considered feasibly informative to stakeholders. Applying this algorithm revealed an optimal value for k of four, but

		Cluster				Population
		0	1	2	3	
Characteristics	**Percentage of spells**	9.90	19.27	69.39	1.44	100.00
	Mean spell cost, £	8051.23	2309.63	1508.41	17888.43	2265.40
	Percentage of recorded costs	29.01	19.38	48.20	3.40	100.00
	Median age	77.00	77.00	71.00	82.00	73.00
	Minimum LOS	12.82	0.01	0.00	48.82	0.00
	Mean LOS	25.31	6.47	4.11	75.36	7.69
	Maximum LOS	51.36	30.86	16.94	224.93	224.93
	Median COPD adm. in last year	2.00	1.00	1.00	2.00	1.00
	Median no. of LTCs	2.00	3.00	1.00	3.00	1.00
	Median no. of ICDs	9.00	8.00	5.00	11.00	6.00
	Median CCI	9.00	20.00	4.00	18.00	4.00
Intervention prevalence	**None, %**	80.19	83.42	65.76	89.74	70.94
	PR, %	15.81	13.43	27.98	8.97	23.69
	SN, %	3.81	2.87	4.63	1.28	4.16
	Both, %	0.19	0.29	1.63	0.00	1.21
LTC prevalence	**Pulmonary disease, %**	100.00	100.00	100.00	100.00	100.00
	Diabetes, %	19.07	28.14	14.84	25.00	17.97
	AMI, %	13.86	22.93	8.76	16.03	12.10
	CHF, %	12.47	53.80	0.00	26.28	11.98
	Renal disease, %	7.53	19.54	1.92	17.95	6.11
	Cancer, %	7.53	12.28	2.93	10.90	5.30
	Dementia, %	6.88	21.26	0.00	26.92	5.17
	CVA, %	8.65	13.33	0.70	19.87	4.20
	PVD, %	4.37	7.69	2.27	5.77	3.57
	CTD, %	5.12	4.25	3.11	4.49	3.55
	Obesity, %	2.51	3.01	1.49	7.69	1.97
	Metastatic cancer, %	1.49	4.54	0.00	0.64	1.03
	Paraplegia, %	1.30	3.73	0.24	0.64	1.02
	Diabetic compl., %	0.19	0.86	0.48	1.92	0.54
	Peptic ulcer, %	1.58	0.81	0.23	1.28	0.49
	Sepsis, %	1.77	0.91	0.15	1.92	0.48
	Liver disease, %	0.28	0.48	0.23	0.00	0.28
	C. diff, %	0.74	0.10	0.01	0.64	0.11
	Severe liver disease, %	0.19	0.43	0.00	0.00	0.10
	MRSA, %	0.28	0.05	0.03	1.28	0.07
	HIV, %	0.00	0.00	0.03	0.00	0.02

Table 1. A summary of clinical and condition-specific characteristics for each cluster and the population

both three and five clusters were considered. Both of these cases were eliminated due to a lack of clear separation in the characteristics of the clusters.

Although the dataset is confidential and may not be published, a synthetic analogue which illustrates the clustering has been archived under . A summary of the dataset and its clustering is provided in Table 1. Note that a negative length of stay indicates that the patient had passed away prior to arriving at the hospital and so these spells have been omitted from further analysis. This table separates each cluster and the overall dataset (referred to as the population). From this table, helpful insights can be gained about the segments identified by the clustering. For instance, the needs of the spells in each cluster can be summarised succinctly:

- Cluster 0 represents those spells with relatively *low clinical complexity but high resource requirements*. The mean spell cost is almost four times the population average, and the shortest spell is almost two weeks long. Moreover, the median number of COPD-related admissions in the last year is elevated, indicating that patients presenting in this way require more interactions with the system.
- Cluster 1, the second-largest segment, represents the spells with *complex clinical profiles despite lower resource requirements*. Specifically, the spells in this cluster have the highest median CCI and number of LTCs, and the highest condition prevalence across all clusters but the second-lowest length of stay and spell costs.

- Cluster 2 represents the majority of spells and those where *resource requirements and clinical complexities are minimal*; these spells have the shortest lengths, and the patients present with fewer diagnoses and a lower median CCI than any other cluster. In addition to this, the spells in Cluster 2 have the highest intervention prevalence. However, they have the lowest condition prevalence across all clusters.
- Cluster 3 represents the smallest section of the population but perhaps the most critical: spells with *high complexity and high resource needs.* The patients within Cluster 3 are the oldest in the population and are some of the most frequently returning despite having the lowest intervention rates. The lengths of stay vary between seven and 32 weeks, and the mean spell cost is almost eight times the population average. This cluster also has the second-highest median CCI, and the highest median number of concurrent diagnoses.

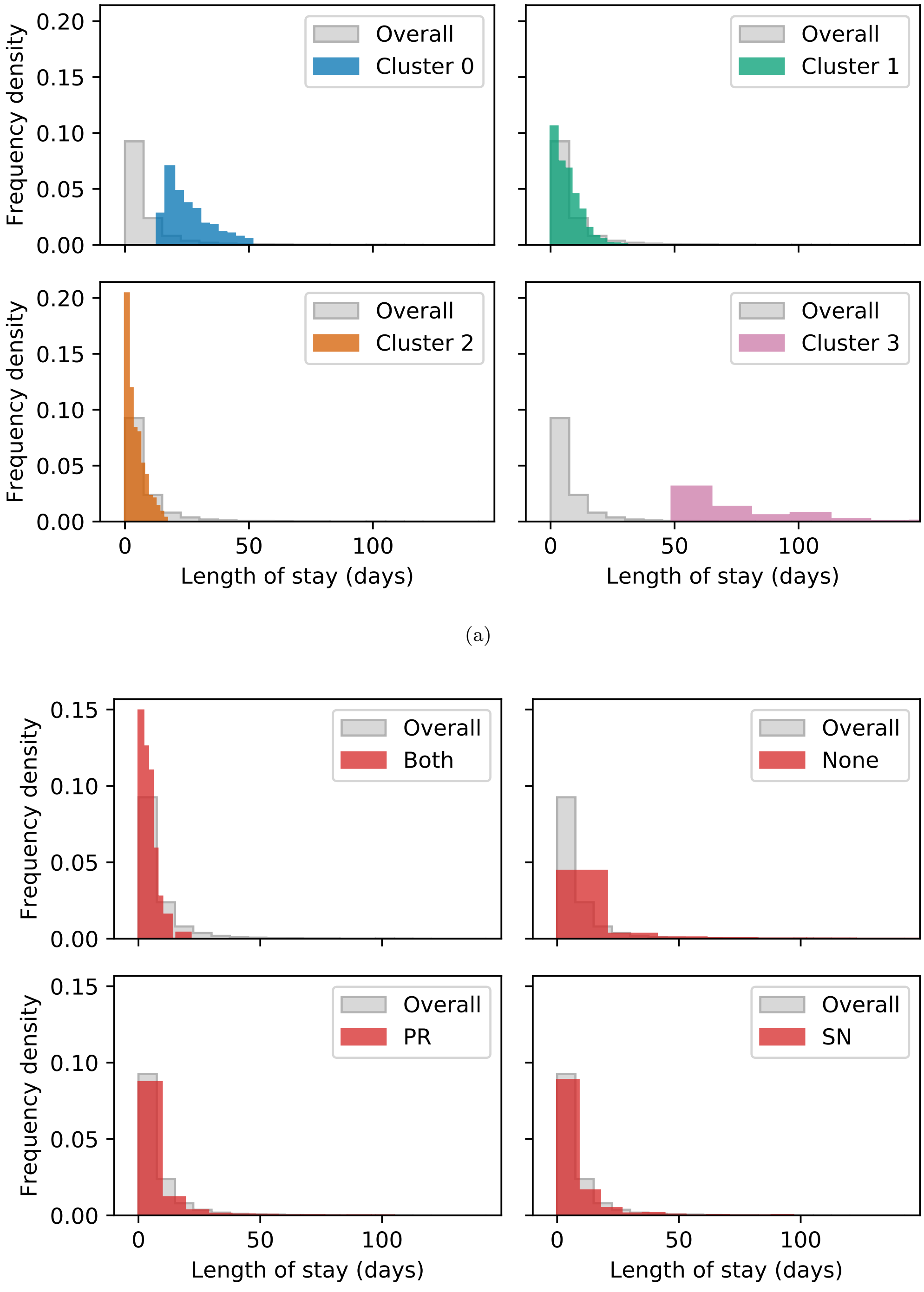
The attributes listed in Table 1 can be studied beyond summaries such as these, however. Figures 2 through 6 show the distributions for some clinical characteristics for each cluster. Each of these figures also shows the distribution of the same attributes when splitting the population by intervention. While this classical approach — of splitting a population based on a condition or treatment — can provide some insight into how the different interventions are used, it has been included to highlight the value added by segmenting the population via data without such a prescriptive framework.

Figure 2 shows the length of stay distributions as histograms. Figure 2a demonstrates the different bed resource requirements well for each cluster — better than Table 1 might — in that the difference between the clusters is not only a matter of varying means and ranges, but entirely different shapes to their respective distributions. Indeed, they are all positively skewed, but there is no real consistency beyond that. When comparing this to Figure 2b, there is undoubtedly some variety, but the overall shapes of the distributions are generally similar. The exception is the spells with no COPD intervention, where binning could not improve the visualisation due to the widespread distribution of their lengths of stay.

The same conclusions can be drawn about spell costs from Figure 3; there are distinct patterns between the clusters in terms of their costs, and they align with the patterns seen in Figure 2. Such patterns are expected given that length of stay is a driving force of healthcare costs. Equally, there does not appear to be any immediately discernible difference in the distribution of costs when splitting by intervention.

Similarly to the previous figures, Figure 4 shows that clustering has revealed distinct patterns in the CCI of the spells within each cluster, whereas splitting by intervention does not. All clusters other than Cluster 2 show clear, heavy tails, and in the cases of Clusters 1 and 3, the body of the data exists far from the origin as indicated in Table 1. In contrast, the plots in Figure 4b all display similar, highly skewed distributions regardless of intervention.

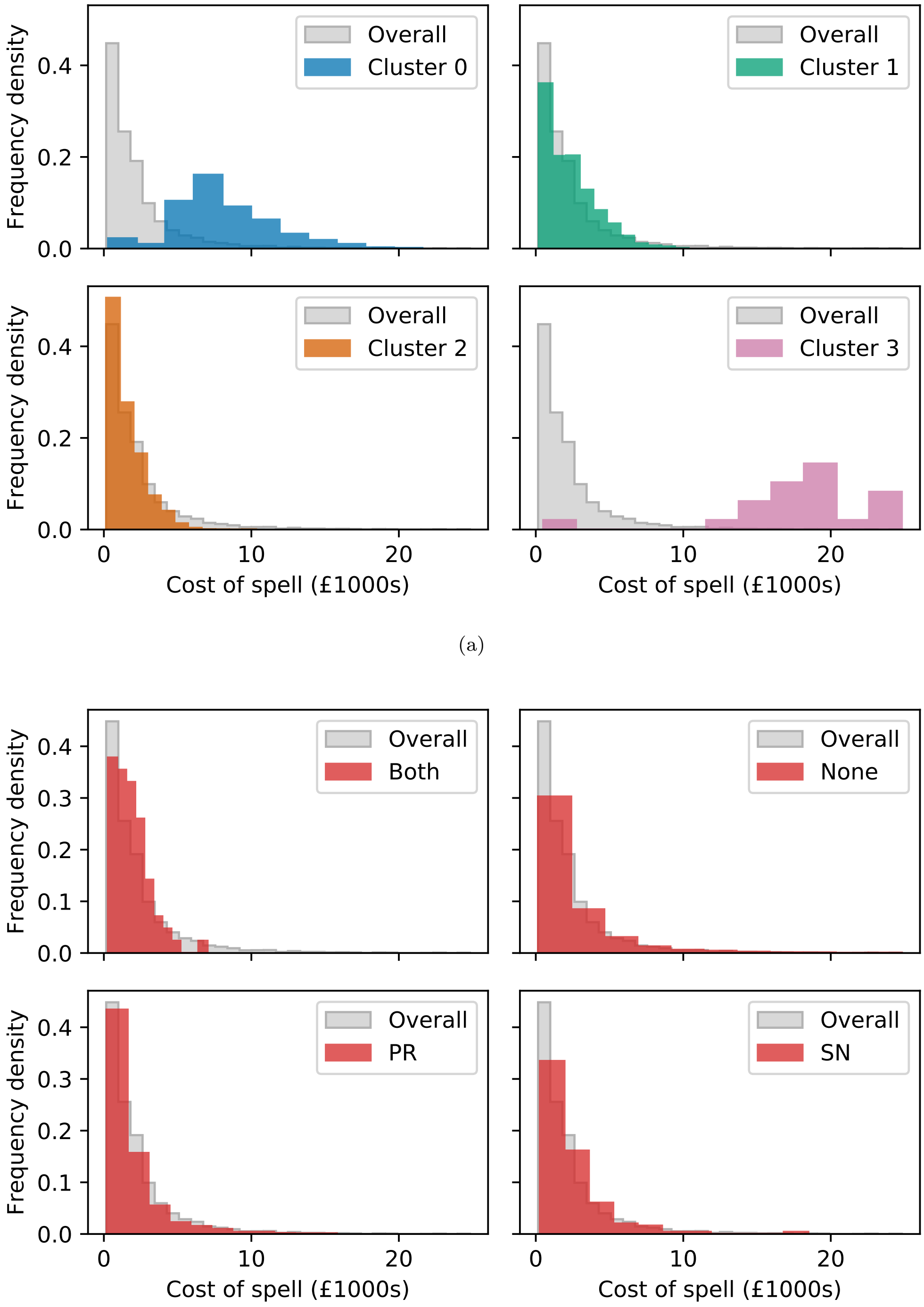
Figures 5 and 6 show the proportions of each grouping presenting levels of concurrent LTCs and ICDs, respectively. By exposing the distribution of these attributes, some notion of the clinical complexity for each cluster can be captured better than with Table 1 alone. In Figure 5a, for instance, there are distinct LTC count profiles among the clusters: Cluster 0 is typical of the population; Cluster 1 shows that no patient presented COPD solely as an LTC in their spells, and more than half presented at least three; Cluster 2 is similar in form to the population but is strongly biased towards patients presenting COPD as the only LTC; Cluster 3 has the closest-to-uniform spread among the four bins despite the increased length of stay and CCI, suggesting a diverse



(a)

(b)

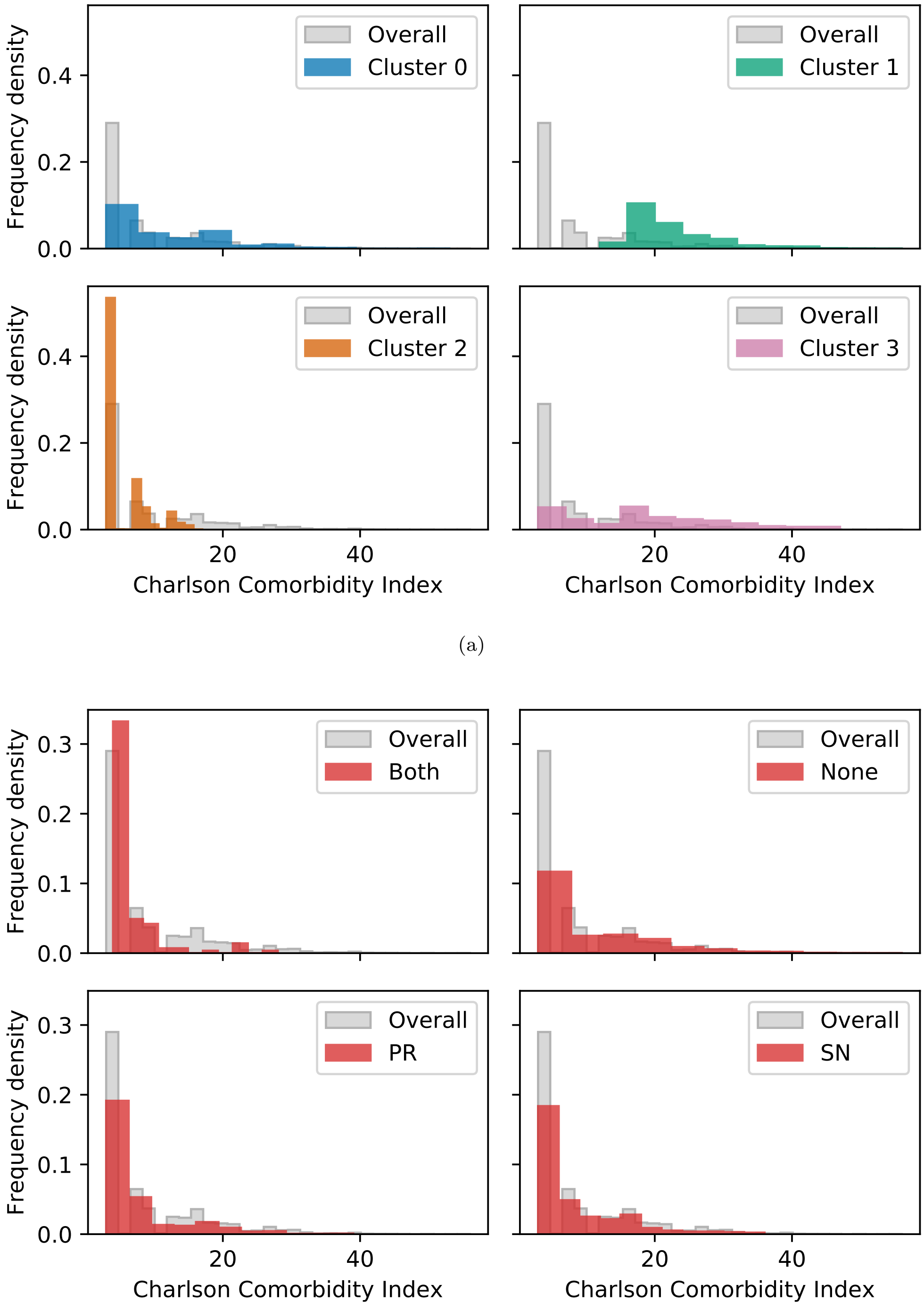
Figure 2. Histograms for length of stay by (a) cluster and (b) intervention



(a)

(b)

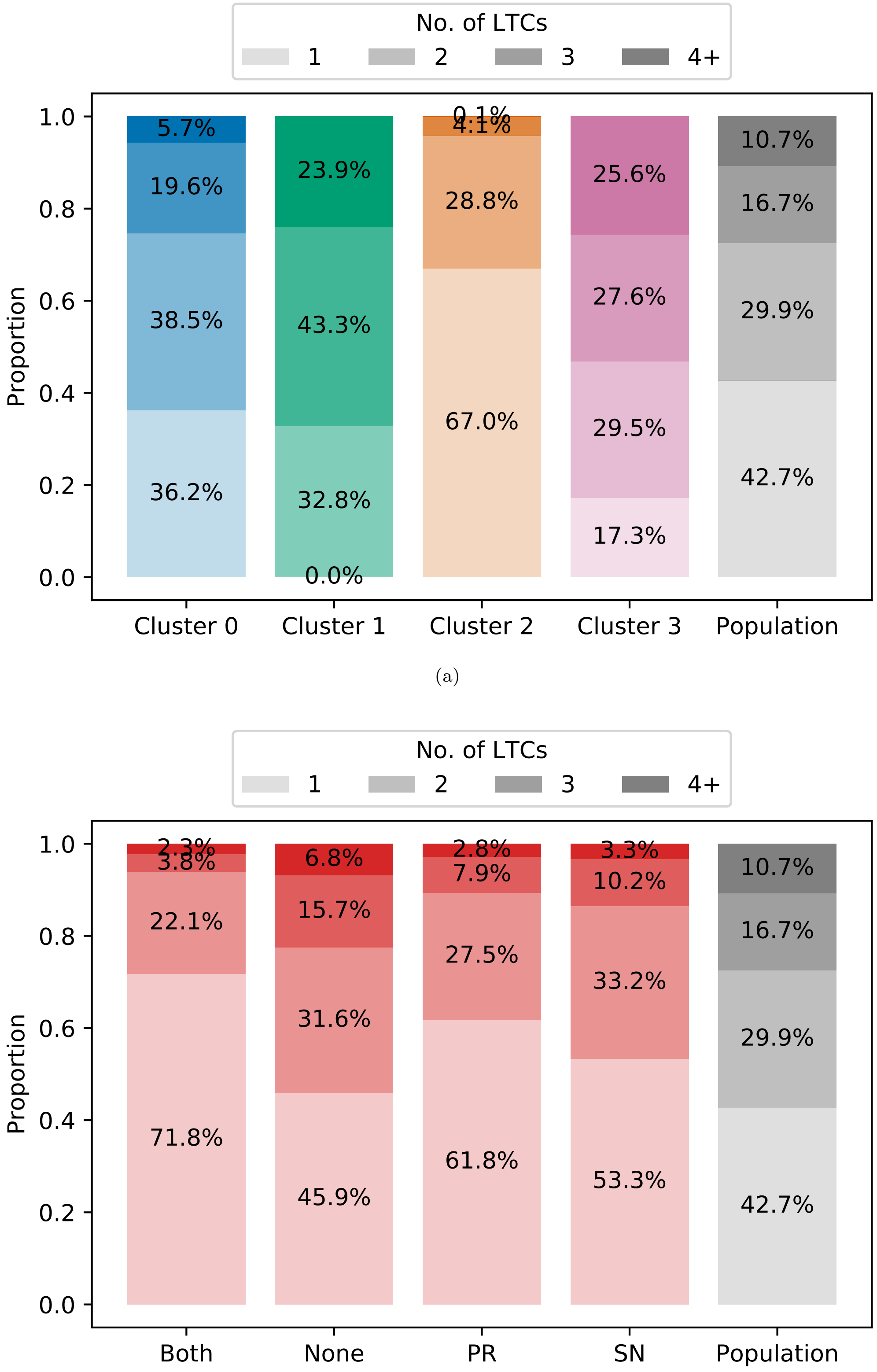
Figure 3. Histograms for spell cost by (a) cluster and (b) intervention



(a)

(b)

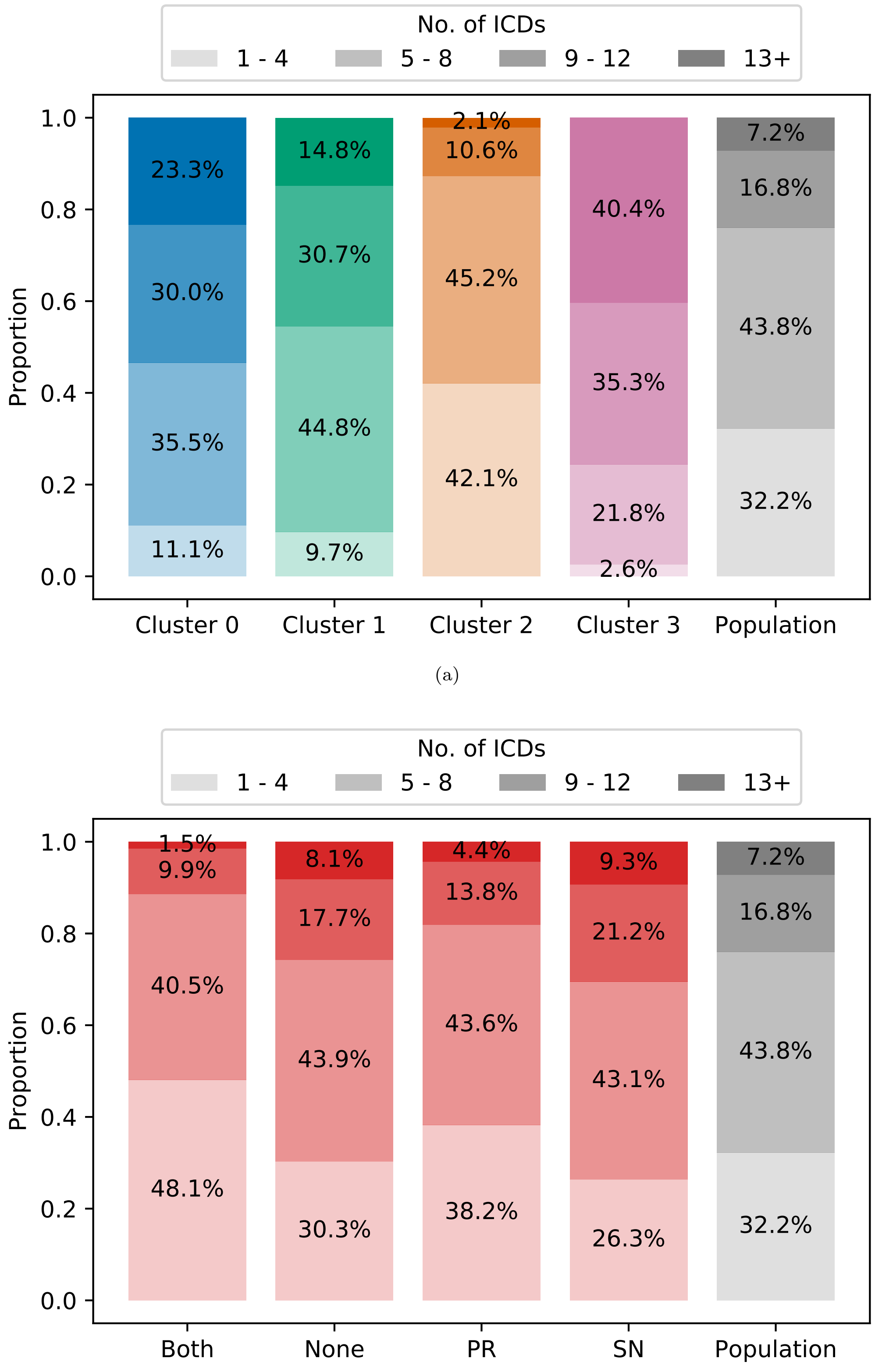
Figure 4. Histograms for CCI by (a) cluster and (b) intervention



(a)

(b)

Figure 5. Proportions of the number of concurrent LTCs in a spell by (a) cluster and (b) intervention



(a)

(b)

Figure 6. Proportions of the number of concurrent ICDs in a spell by (a) cluster and (b) intervention

array of patients in terms of their long term medical needs.

Figure 6a largely mirrors these cluster profiles with the number of concurrent ICDs. There are some points of interest, however. Firstly, Cluster 1 has a relatively low-leaning distribution of ICDs that does not marry up with the high rates of LTCs. Secondly, the vast majority of spells in Cluster 3 present with at least nine ICDs suggesting a likely wide range of conditions and comorbidities beyond the LTCs used to calculate CCI.

However, little can be drawn from the intervention counterparts to these figures (i.e. Figures 5b and 6b), regarding the corresponding spells. One thing of note is that patients receiving both interventions for their COPD (or either, in fact) have disproportionately fewer LTCs and concurrent ICDs when compared to the population. Aside from this, the profiles of each intervention are similar to one another.

As discussed earlier, the purpose of this manuscript is to construct a queuing model for the data described here. Insights have already been gained into the needs of the segments that have been identified in this section. However, to glean further insights, some parameters of the queuing model must be recovered from the data. The following two sections briefly introduce queues, and describe how these parameters are derived using the dataset at hand, respectively.

2. An introduction to queues

Queues facilitate the orderly provision of services. Examples include lining up to board a bus, assembly lines in a factory, or patients arriving at a hospital. In all queues there are two types of agent: those providing the service (a bus driver), and those demanding it (the passengers). The generic terminology for these agents are *servers* and *customers*, respectively.

As well as individual queues, networks of interconnecting queues can be described in a similar manner. In a *queuing network*, each individual queue is called a *node*. For instance, a hospital could be considered a network of queues, where patients arrive into triage, are processed, and are redirected throughout their spell.

The observed characteristics of a queuing system can be used to construct a mathematical model. Such a model would describe things like the process by which customers arrive to a queue, the rules that allow customers to be served, and the time taken to serve a customer. Queuing theory is the branch of mathematics concerned with the analysis of these models. Queuing theory is a mature discipline with many facets that extend beyond the needs of this manuscript. Comprehensive and informative introductions to queuing models, queuing theory, and the simulating of queues can be found in Bhat (2015); Shortle, Thompson, Gross, and Harris (2018); Stewart (2009). Further, applications of queuing models to healthcare systems are plentiful, but examples include Bittencourt, Verter, and Yalovsky (2018); Cochran and Roche (2009); Mohammadi and Salehi-Rad (2012); Steins and Walther (2013); Williams et al. (2015); Yom-Tov and Mandelbaum (2014).

2.1. Elements of a queue

A queue is made up several components: the service facility, a number of servers within that facility, a line in which customers wait to be served, and a stream of arriving customers. Figure 7 shows a diagram of a queue. The characteristics associated with the components of a queue are often summarised using Kendall's notation Stewart (2009).

The exact notation varies somewhat, but here it shall be denoted $A/S/c/m/K/Q$. This notation also defines the *parameters* of the queue. The process in Section 3 estimates some unknown parameters of a queue.

Kendall's notation acts as a shorthand to fully describe a queue, and is as follows. Customers arrive to the queue according to an *inter-arrival time distribution*, A, and wait in line to be served according to a *queuing discipline*, Q. Typically, customers are served as they arrive. This discipline is called FIFO (first in first out). Other disciplines include LIFO (last in first out) and priority scheduling — as used in emergency triage. At the service facility there are c parallel servers, who each serve customers according to the *service time distribution*, S. Sometimes it is beneficial to attach a capacity to the system, denoted by $K \geq c$. If omitted, an unlimited system capacity is presumed. The *system capacity* can also be distinguished from an optional *queue capacity*, $m < K$, which limits the number of customers allowed to wait in line. Again, this is assumed to be ∞, unless specified.

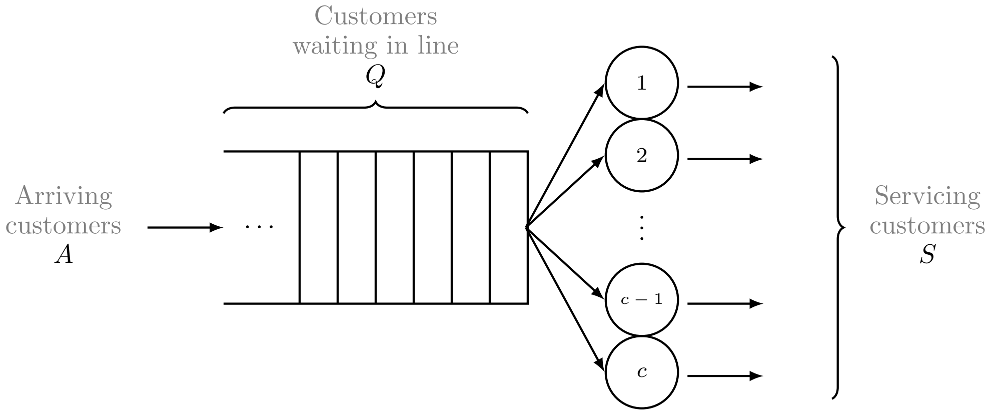


Figure 7. The anatomy of a queue

The distributions used to model inter-arrival and service times are numerous, but some commonly studied examples are:

- Markovian (denoted M). Customers arrive according to a Poisson process. The inter-arrival or service times follow an exponential distribution with rate $\alpha > 0$, i.e. they have probability density function:

$$f(t) = \alpha e^{-\alpha t}; \quad t \geq 0 \tag{3}$$

- Deterministic (denoted D). Inter-arrival or service times are non-stochastic and are of fixed length.
- General (denoted G). Arrivals are random, and inter-arrival or service times follow a general probability distribution.

2.2. Some classical queues

2.2.1. The $M/M/1$ queue

One of the best-known queues is the $M/M/1$ queue. In this queue, customers arrive according to a Poisson process at a rate of λ. The customers are served by a single server exponentially, at a rate of μ. Owing to the memoryless property of the exponential distribution, this queue can be represented as a continuous-time Markov chain over the state space $\mathbb{N}_0 = \{0, 1, 2, \dots\}$.

A stochastic process, (X_t), which is defined over a countable state space, $\mathcal{S}$, is considered a *Markov chain* if and only if for all $n \in \mathbb{N}$ and for any $(n+1)$-tuple of states, $s_0, \dots, s_n \in \mathcal{S}^n$, the process satisfies:

$$\mathbb{P}\left(X_n = s_n \mid X_{n-1} = s_{n-1}, \dots, X_0 = s_0\right) = \mathbb{P}\left(X_n = s_n \mid X_{n-1} = s_{n-1}\right) \tag{4}$$

That is, the probability of being in state s_n is dependent only on the previous state, s_{n-1}. The Markov chain underlying an $M/M/1$ queue is also known as a *birth-death process*. Figure 8 shows a diagram of the birth-death process of an $M/M/1$ queue.

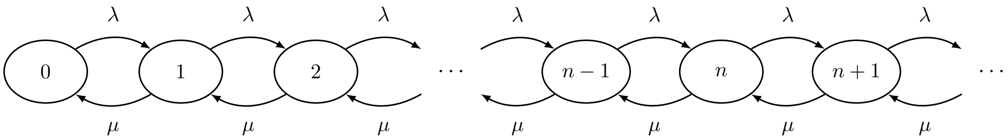


Figure 8. A state space diagram for an $M/M/1$ queue

An important property of the $M/M/1$ queue is its *traffic intensity*, which is defined as $\rho = \frac{\lambda}{\mu}$. This quantity also represents the proportion of time that the server spends serving customers, and so measures their *utilisation*. The $M/M/1$ queue is considered *stable* (i.e. the underlying process will become stationary eventually) if $\rho < 1$. In an unstable queue, customers arrive at a faster rate than they are served, and so the line grows indefinitely.

Many other properties of the $M/M/1$ queue can be explicitly derived from this representation, including steady-state solutions, the expected size of the system, and average response times. The calculation of these quantities also makes use of Little's Law **?**, which relates average system size, L, with average waiting time, W, in stationary processes such that:

$$L = \lambda W \tag{5}$$

2.2.2. The M/M/c queue

The $M/M/c$ queue is an extension of the $M/M/1$ queue where there are $c \in \mathbb{N}$ independent servers working in parallel. Customers still arrive according to a Poisson process with rate λ, and customer service times follow an exponential distribution with a mean of $\frac{1}{\mu}$. The traffic intensity of the $M/M/c$ queue is also $\rho = \frac{\lambda}{\mu}$, but server utilisation is measured as the mean traffic intensity across the servers, i.e. $\frac{\rho}{c}$. The stability condition for the $M/M/c$ queue is $\rho < c$.

As with the $M/M/1$ queue, the $M/M/c$ queue can be represented as a birth-death process on the state space $\mathbb{N}_0$, as shown in Figure 9. Since there are multiple servers, some servers may be idle when there are customers in the system. Once all servers are active, arriving customers join the line and wait for service.

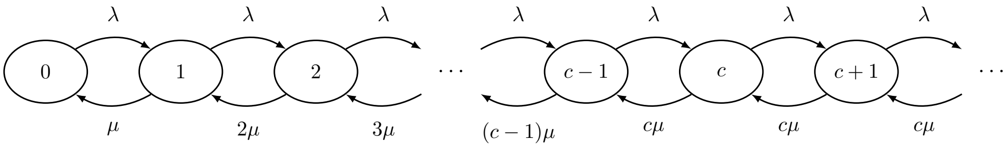


Figure 9. A state space diagram for an $M/M/c$ queue

2.2.3. The $M/G/c$ queue

As useful as memoryless service times are in deriving properties of queues, they are not always representative of real systems. The $M/G/c$ queue extends the $M/M/c$ queue to allow for a general service time distribution. Again, there are c parallel servers, and customers arrive randomly at a rate of λ.

The $M/G/c$ queue cannot be represented as a Markov chain, but it is a stochastic process on the same state space as the other queues in this section. Given the generic nature of customer departure times, a lot of the structure of the underlying process is lost. As such, deriving exact values for many properties of the $M/G/c$ queue continues to be an open problem Kingman (2009). Despite the theoretical challenges posed by the $M/G/c$ queue, simulating these generic queues can still be of great benefit — as is done in Section 3.

2.3. Simulation tools

As well as theoretical results, queues provide a valuable basis for computer simulation. Theoretical models are limited when studying complex queuing systems, where the parameterisation of a system requires complicated notation or derivations to produce useful results. When properly utilising computer simulation, the stochastic intricacies of a system may be more readily observed. Examples of such systems include the multi-class queuing networks studied in Cochran and Roche (2009).

There are numerous tools available for simulating queues, but many leave the associated research prone to issues like reproducibility — as mentioned in Section **??**. A recent review on the subject and its tools is Dagkakis and Heavey (2016). One of the defining features of a simulation tool is whether it has a *graphical user interface* (GUI) or not. GUIs provide accessibility to the non-technical members of a simulation project, but can also foster poor simulation practices Bell and O'Keefe (1987).

The simulation framework of choice for this manuscript is the discrete event simulation library, Ciw Palmer et al. (2019). Ciw is written in Python, and is a well-developed piece of open-source software, adhering to best practices in research software development. In Palmer et al. (2019), the authors stress how ensuring sustainable and reproducible simulation work are at the core of their development process.

3. Constructing the queuing model

The data available for study in this work is not as detailed as in comparative projects. Without access to such data — but intending to gain insight from what is available — it is imperative to bridge the gap left by the incomplete data. Figure 10 provides a diagrammatic depiction of the process described in this section.

It is often the case that in practical situations where suitable data is not (immediately) available, further inquiry in that line of research will stop. Queuing models in healthcare settings appear to be such a case; the line ends at incomplete queue data. The bibliographic work Asanjarani et al. (2017) collates articles on the estimation of queuing system characteristics — including their parameters. Despite its breadth of almost 300 publications from 1955, only two articles have been identified as being applied to healthcare: Mohammadi and Salehi-Rad (2012); Yom-Tov and Mandelbaum (2014). Both works are concerned with customers who can re-enter services during their time in the queuing system, which is mainly of value when considering the effect of unpredictable behaviour in intensive care units, for instance. In Mohammadi and Salehi-Rad (2012), the authors seek to approximate service and re-service densities through a Bayesian approach and by filtering out those customers seeking to be serviced again. Meanwhile, the approach in Yom-Tov and Mandelbaum (2014) considers an extension to the $M/M/c$ queue with direct re-entries. The devised model is then used to determine resource requirements in two healthcare settings.

Aside from healthcare-specific works, the approximation of queue parameters has formed a part of relevant modern queuing research. However, the scope is primarily focused on theoretic approximations rather than simulation. For instance, two recent works Djabali et al. (2018); Goldenshluger (2016) consider an underlying process to estimate a general service time distribution in single server and infinite server queues respectively.

While these solutions are interesting, they do not necessarily tackle the issue in this scenario where information about the system is also missing. With that, there is a precedent for simplifying healthcare systems to a single node with parallel servers that emulate overall resource availability. Two studies Steins and Walther (2013); Williams et al. (2015) provide 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, the authors of Williams et al. (2015) show 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.

3.1. Deriving the model parameters

Following in the suit of recent literature Steins and Walther (2013); Williams et al. (2015), this work employs a single node using the $M/M/c$ queue to model a hypothetical ward of patients presenting COPD. In addition to this, the grouping found in Section 1.2 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 distribution is used to simplify the model parameterisation.
(2) There are $c \in \mathbb{N}$ servers available to arriving patients at the node representing the overall resource availability, including bed capacity and hospital staff.
(3) There is no queue or system capacity. In Williams et al. (2015), 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 the sole unit for COPD patients within the health board, this assumption is not held.
(4) Without the availability of expert clinical knowledge, a first-in-first-out service policy is employed in place of some patient priority framework.

Each group of patients has its arrival distribution, the parameter of which is the

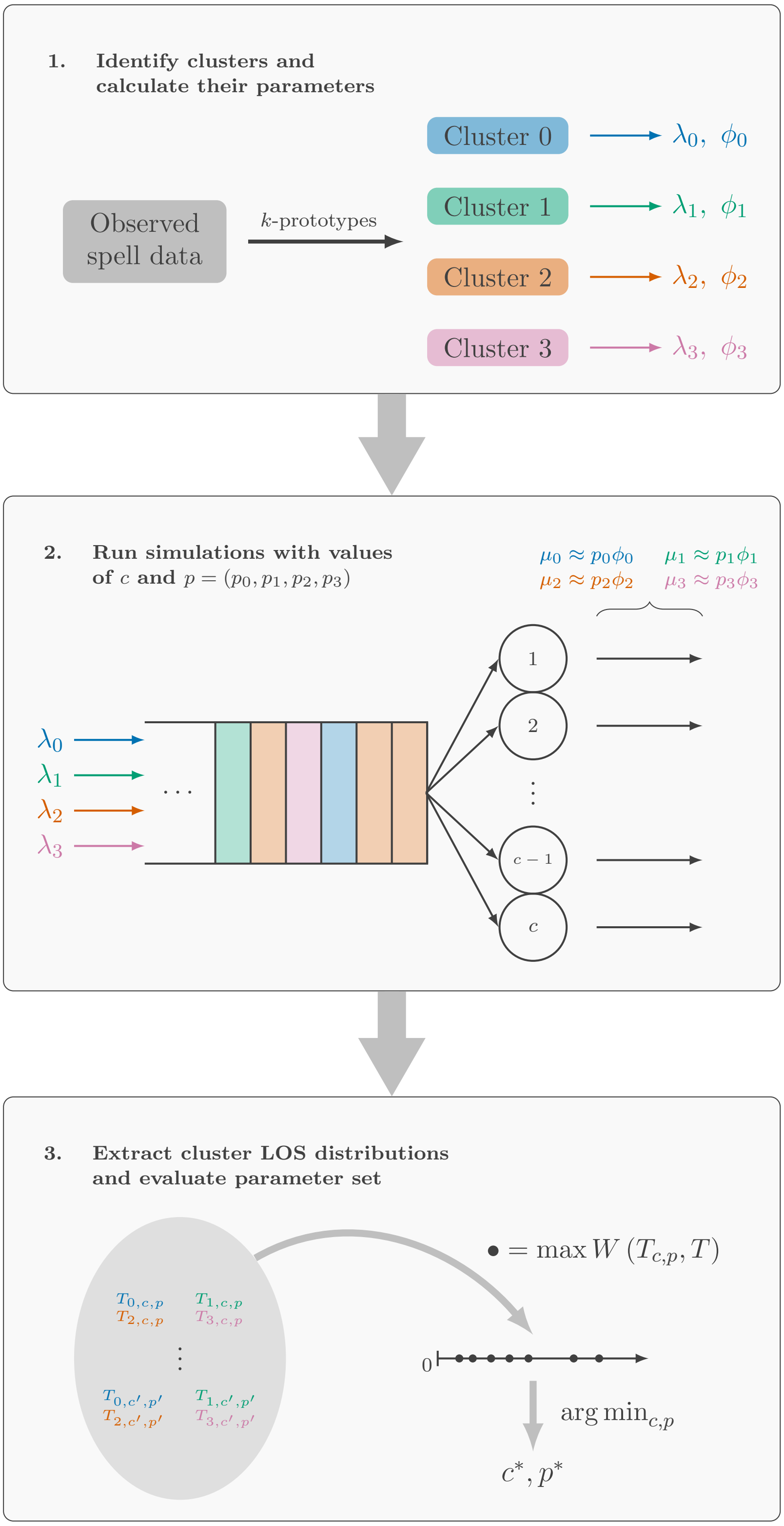


Figure 10. A diagrammatic depiction of the queuing parameter recovery process

reciprocal of the mean inter-arrival times for that group. This parameter is denoted by λ_l for each cluster l.

Like arrivals, each group of patients has its service time distribution. Without full details of the process order or idle periods during a spell, some assumption must be made about the actual 'service' time of a patient in the 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. As indicated by the distributions in Figure 2a, the length of stay distributions require shifting prior to fitting an exponential distribution.

Let T_l denote the set of observed lengths of stay for cluster l, and let $m_l = \max\{0, \min T_l\}$ be its feasible minimum. Thus, the *shifted times* for cluster l, denoted $\widehat{T_l}$, are:

$$\widehat{T_l} := \{t - m_l : t \in T_l\} \tag{6}$$

An exponential distribution may be fitted to these shifted system times by using their mean, denoted by $\frac{1}{\phi_l}$, as the distribution parameter. For the sake of simplicity, it is assumed that for each cluster l, the mean shifted service time of that cluster, $\frac{1}{\mu_l}$, is proportional to the corresponding mean shifted system time such that:

$$\mu_l = p_l \phi_l \tag{7}$$

where $p_l \in (0, 1]$ is a *service proportion* parameter to be determined for each group.

With these definitions, the service time for cluster l, denoted S_l, is distributed by a *shifted exponential distribution* with a mean of $\frac{1}{\mu_l}$ and shift of m_l. The probability density function of this distribution is as follows:

$$f(s) = \begin{cases} \mu_l e^{-\mu_l(s-m_l)} & \text{if } s \geq m_l \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Since this distribution is geometrically identical to the exponential distribution with rate μ_l except for a shift of m_l, its memoryless property holds for $s \geq m_l$. However, since this model allows for multiple classes and the shift terms are not the same for each cluster, this model technically should be reclassified as a $M/G/c$ model. Regardless of this, the mean service time for spells in cluster l is given by:

$$\mathbb{E}(S_l) = \int_{m_l}^{\infty} \mu_l s e^{-\mu_l(s-m_l)} \mathrm{d}s = m_l + \frac{1}{\mu_l} \tag{9}$$

3.2. Validating the model

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

Several methods are available for the statistical comparison of two or more distributions, such as the Kolmogorov-Smirnov test, a variety of discrepancy approaches such as summed mean-squared error, and f-divergences. A popular choice among the last group (which may be considered distance-like) is the Kullback-Leibler divergence which measures relative information entropy from one probability distribution to another Kullback and Leibler (1951). A key issue with many of these methods is that they lack interpretability, something which is paramount when conveying information to stakeholders, not only for explaining how something works, but also how its results may be explained.

As such, a reasonable candidate is the (first) Wasserstein metric, also known as the 'earth mover' or 'digger' distance Vaserstein (1969). The Wasserstein metric satisfies the conditions of a formal mathematical metric — like Euclidean distance or the dissimilarity measure given in Definition **??**. Also, the values of the Wasserstein metric take the units of the distributions under comparison (in this case: days). These characteristics can aid understanding and explanation. 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 moves and the distance by which it must be moved. More formally, the Wasserstein distance between two probability distributions U and V is defined as:

$$W(U,V) = \int_0^1 \left|F^{-1}(t) - G^{-1}(t)\right| dt \tag{10}$$

Here, F and G are the cumulative density functions of U and V, respectively. A proof of (10) is presented in Ramdas, Trillos, and Cuturi (2017).

Each trial used here takes a parameter set and simulates the ward across a series of independent repetitions. The parameter set with the smallest maximum distance between the simulated system time distribution and the observed length of stay distribution is taken to be the most appropriate. To be specific, let $T_{c,p}$ denote the system time distribution obtained from a simulation with c servers and $p := (p_0, p_1, p_2, p_3)$, and let T denote the observed length of stay distribution. Then the optimal parameter set (c^*, p^*) is given by:

$$(c^*, p^*) = \underset{c,p}{\arg\min} \left\{\max\left\{W\left(T_{c,p}, T\right)\right\}\right\} \tag{11}$$

The parameter sweep included values of each p_l from 0.5 to 1.0 with a granularity of 5.0×10^{-2} and values of c from 30 to 50 at steps of five. 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 are summarised in Figures 11 through 13. Each plot shows a comparison of the observed lengths of stay across all groups and the newly simulated data with the best, median and worst parameter sets, respectively. These figures highlight the importance of choosing good parameters under this model as the differences in the quality of the fits are stark. In the best case the fit is uncanny,

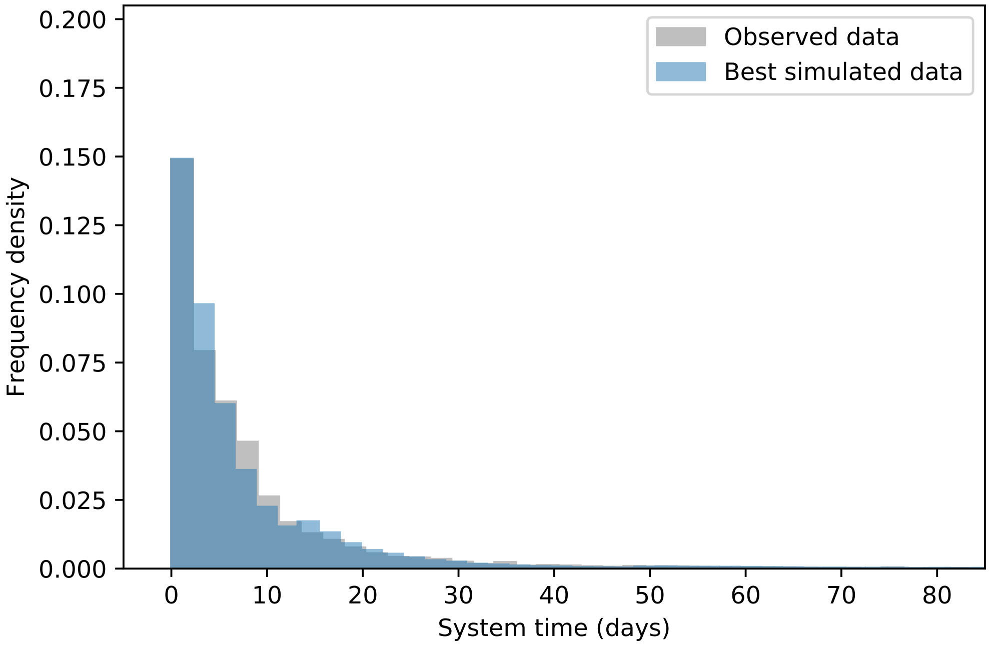


Figure 11. Histograms of the best-simulated and observed LOS data

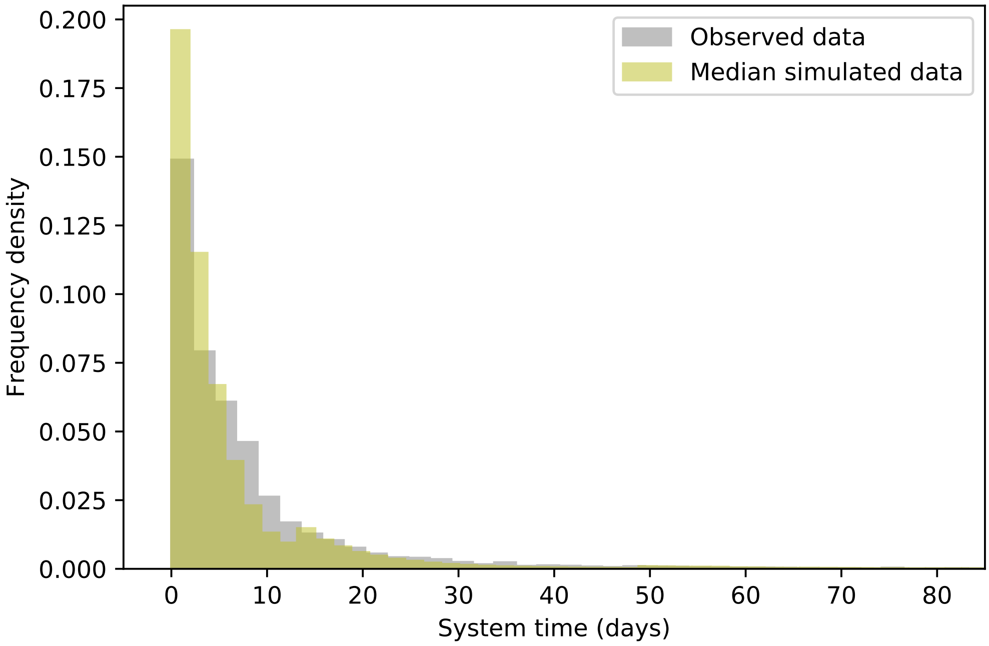


Figure 12. Histograms of the median-simulated and observed LOS data

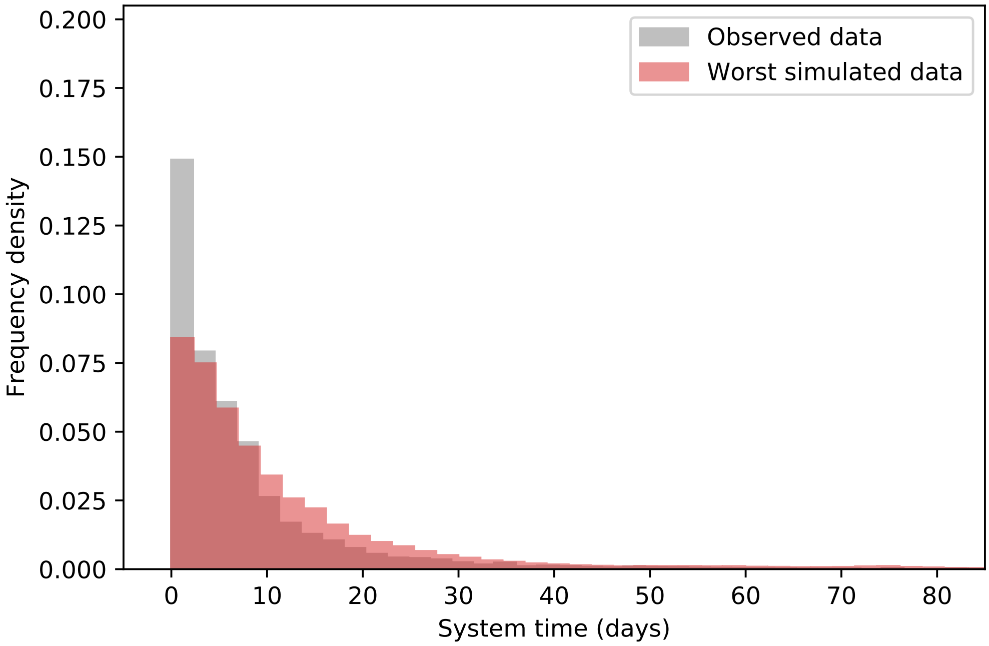


Figure 13. Histograms of the worst-simulated and observed LOS data

whereas the median case shows a distribution that inflates the presence of short-stay patients despite an otherwise good fit. Meanwhile, Figure 13 displays a distribution that only resembles the observed distribution in its positive skew; the worst-case distribution lacks the distinctive 'exponential' nose and has a considerably heavier tail corresponding to a disproportionate amount of long-stay patients. Table 2 reinforces these results numerically, showing a precise fit by the best parameter set across all measures, except the maximum recorded stay.

In this section, the previously identified clustering enriched the overall queuing model and was used to recover the parameters for several classes within that. Now, using this model, the next section details an investigation into the underlying system by adjusting the parameters of the queue with the clustering.

4. Adjusting the queuing model

This section comprises several what-if scenarios — a classic component of healthcare operational research — under the novel parameterisation of the queue established in Section 3. The outcomes of interest in this work are server (resource) utilisation and system times. These metrics capture the driving forces of cost and the state of the system. Specifically, the objective of these experiments is to address the following questions:

- How is the system affected by a change in overall patient arrivals?
- How is the system affected by a change in resource availability?
- How is the system affected by patients moving between clusters?

Given the nature of the observed data, the queuing model parameterisation and its

		Observed	Best simulated	Median simulated	Worst simulated
Model characteristic	p_0	NaN	0.80	0.70	1.00
	p_1	NaN	1.00	0.55	1.00
	p_2	NaN	1.00	0.85	0.95
	p_3	NaN	0.85	0.70	0.90
	c	NaN	35.00	40.00	30.00
	Max. distance	0.00	0.68	1.95	44.25
LOS statistic	**Mean**	7.70	7.56	6.23	11.56
	Std.	11.86	11.44	10.45	14.81
	Min.	0.00	0.00	0.00	0.00
	25%	1.49	1.60	1.16	3.00
	Med.	4.20	3.90	2.90	6.90
	75%	8.93	8.81	6.54	14.21
	Max.	224.93	219.92	187.78	230.49

Table 2. A comparison of the observed and simulated data based on the model parameters and summary statistics for length of stay

assumptions, the effects on the chosen metrics in each scenario are in relative terms with respect to the base case. The base case being those results generated from the best parameter set recorded in Table 2. In particular, the data from each scenario is scaled by the corresponding median value in the base case, meaning that a metric having a value of 1 is 'normal'.

As mentioned in Section **??**, the source code used throughout this manuscript has been archived online under . Also, the datasets generated from the simulations in this section, and the parameter sweep, have been archived online .

4.1. Changes to overall patient arrivals

Changes in overall patient arrivals to a queue reflect real-world scenarios where some stimulus is improving (or worsening) the condition of the patient population. Examples of stimuli could include an ageing population or independent life events that lead to a change in deprivation, such as an accident or job loss. Within this model, overall patient arrivals are altered using a scaling factor denoted by $\sigma > 0$. This scaling factor is applied to the model by multiplying each cluster's arrival rate by σ. That is, the new arrival rate for a cluster, l, denoted $\hat{\lambda}_l$, is given by:

$$\hat{\lambda}_l = \sigma \lambda_l \tag{12}$$

Figure **??** shows the effects of changing patient arrivals on (**??**) relative system times and (**??**) relative server utilisation for values of σ from 0.5 to 2.0 at a precision of 1.0×10^{-2}. Specifically, each plot in the figure (and the subsequent figures in this section) shows the median and interquartile range (IQR) of each relative attribute. These metrics provide an insight into the experience of a typical user (or server) in the system. Furthermore, they reveal the stability and variation of the body of users (or servers).

What is evident from these plots is that things are happening as one might expect: as arrivals increase, the strain on the system increases. However, it should be noted that it also appears that the model has some amount of slack relative to the base case. Looking at Figure **??**, for instance, the relative system time distribution stays unchanged up to $\sigma \approx 1.2$, or an approximate 20% increase in arrivals of COPD patients. Beyond that, relative system times quickly rise to an untenable point where the median time becomes orders of magnitude above the norm.

However, Figure **??** shows that the situation for the system's resources reaches its worst-case near to the start of that spike in relative system times (at $\sigma \approx 1.3$). That is, the median server utilisation reaches a maximum (this corresponds to constant utilisation) at this point, and the variation in server utilisation disappears entirely. The reality of this situation is that the system has no slack at all, and all parts of the system are under equal load, which is not preferable given the differences in resource requirements for the parts of a hospital system. For instance, if surgical theatres were in constant use but administrative processing required an equivalent amount of resources to continue running, the system would likely falter or deteriorate entirely.

4.2. Changes to resource availability

As is discussed in Section 3, the resource availability of the system is captured by the number of parallel servers, c. Therefore, to modify the overall resource availability, only the number of servers needs to be changed. This kind of sensitivity analysis is usually done to determine the opportunity cost of adding service capacity to a system, e.g. would an increase of n servers sufficiently increase efficiency without exceeding a budget?

To reiterate the beginning of this section: all suitable parameters are given in relative terms, including the number of servers here. By doing this, the changes in resource availability are more evident, and do away with any concerns as to what a particular number of servers precisely reflects in the real world, be it any combination of hospital beds, equipment availability and medical staff.

Figure **??** shows how the relative resource availability affects relative system times and server utilisation. In this scenario, the relative number of servers took values from 0.5 to 2.0 at an equivalent step size of one in the number of servers, i.e. c takes values from 17 to 70. Overall, these figures fortify the claim from the previous scenario that there is some room to manoeuvre so that the system runs 'as normal', but pressing on those boundaries results in massive changes to both resource requirements and system times.

In Figure **??** this amounts to a maximum of 10% slack in resources before relative system times are substantially affected; further reductions quickly result in a potentially tenfold increase in the median system time, and up to 100 times once resource availability falls by 50%. Moreover, the variation in the body of the relative times (i.e. the IQR) decreases as resource availability decreases. The reality of this is that patients arriving at a hospital are forced to consume more significant amounts of resources (by merely being in a hospital) regardless of their condition, putting added strains on the system. Figure **??** mirrors these observations on the small amount of slack in resource requirements, but (as with the previous scenario) constant utilisation occurs quickly.

Meanwhile, it appears that there is no tangible change in relative system times given an increase in the number of servers. This indicates that the model carries sufficient resources to cater to the population under normal circumstances and that adding service capacity will not necessarily improve system times.

Again, Figure **??** shows that there is a substantial change in the variation in the relative utilisation of the servers. In this case, the variation dissipates as resource levels fall, and increases as resources increase. While the relationship between real hospital resources and the number of servers is not exact, having variation in server utilisation would suggest that small parts of an existing system may be configured or partitioned away in the case of some significant public health event (such as a global pandemic) without overloading the system.

4.3. Moving arrivals between clusters

This scenario is perhaps the most relevant to actionable public health research of those presented here. The clusters identified in this work could be characterised by their clinical complexities and resource requirements, as done in Section 1.2. Therefore, being able to model the movement of some proportion of patient spells from one cluster to another will reveal how those complexities and requirements affect the system itself. The reality is then that if some public health policy could be implemented to initiate that movement informed by a model such as this, then change would be seen in the real system.

In order to model the effects of spells moving between two clusters, the assumption is that each cluster's service time distribution stays the same (and so does each cluster's p_l), but their arrival rates are altered according to some transfer proportion. Consider two clusters indexed at l and m, and their respective arrival rates, λ_l, λ_m. Let $\delta \in [0, 1)$ denote the proportion of arrivals to be moved from cluster l to cluster m. Then the new arrival rates for each cluster, denoted by $\hat{\lambda}_l, \hat{\lambda}_m$ respectively, are:

$$\hat{\lambda}_l = (1 - \delta)\,\lambda_l \quad \text{and} \quad \hat{\lambda}_m = \delta\lambda_l + \lambda_m \tag{13}$$

By moving patient arrivals between clusters in this way, the overall arrivals are left the same since the sum of the arrival rates is the same. Hence, the (relative) effect on server utilisation and system time can be measured independently.

Figures 14 and 15 show the effect on relative system time and relative server utilisation, respectively, of moving patient arrivals between clusters. In each figure, the median and IQR for the corresponding attribute is shown, as in the previous scenarios. Each scenario was simulated using values of δ from 0.0 to 0.98 at steps of 2.0×10^{-2}.

Considering Figure 14, it appears that each type of transfer falls into one of two categories: either completely derailing the system (such as moving any cluster to Cluster 3) or improving system times, albeit mildly. The latter case occurs in the following transfers:

- Cluster 0 to Clusters 1 or 2
- Cluster 1 to Cluster 2
- Cluster 3 to any other cluster

A finer look at the effect of these transfer types on relative system times is given in Table 3. Likewise, their effects on relative server utilisation is given in Table 4.

The message delivered by these transfers is that in order to improve system times in hospitals, the only solution is for the patients arriving at hospital to present with fewer resource requirements. Meanwhile, the complexity of their condition is less influential. Achieving such reductions in resource requirements is certainly no mean feat, but could be addressed by investing in more advanced medical infrastructure in other parts of the healthcare system, beyond hospitals. Furthermore, this could be achieved by implementing some preventive policy that would help improve the overall health of the COPD population, with particular targeting for those most-affected by the

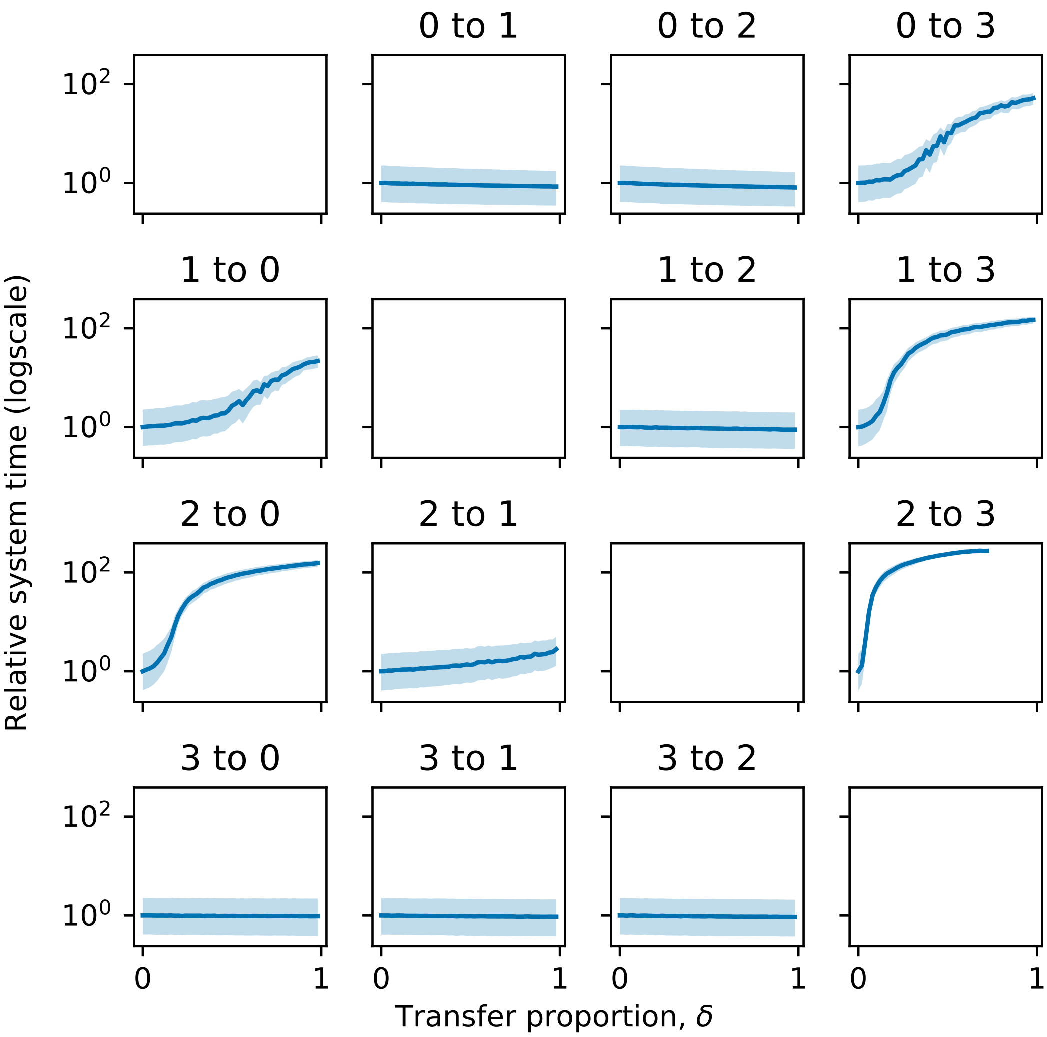


Figure 14. Plots of proportions of each cluster moving to another against relative system time

	δ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Origin	**Destination**										
0	**1**	0.0	-0.0251	-0.0511	-0.0657	-0.0794	-0.0944	-0.1117	-0.1230	-0.1357	-0.1484
	2	0.0	-0.0287	-0.0556	-0.0841	-0.1034	-0.1214	-0.1354	-0.1527	-0.1663	-0.1789
1	**2**	0.0	-0.0048	-0.0072	-0.0393	-0.0452	-0.0606	-0.0762	-0.0761	-0.0909	-0.1058
3	**0**	0.0	-0.0024	-0.0066	-0.0111	-0.0102	-0.0186	-0.0292	-0.0333	-0.0292	-0.0325
	1	0.0	-0.0021	-0.0156	-0.0229	-0.0257	-0.0327	-0.0443	-0.0486	-0.0521	-0.0583
	2	0.0	-0.0182	-0.0242	-0.0298	-0.0365	-0.0337	-0.0487	-0.0554	-0.0530	-0.0646

Table 3. Proportional changes in median relative system time for selected cluster transfers

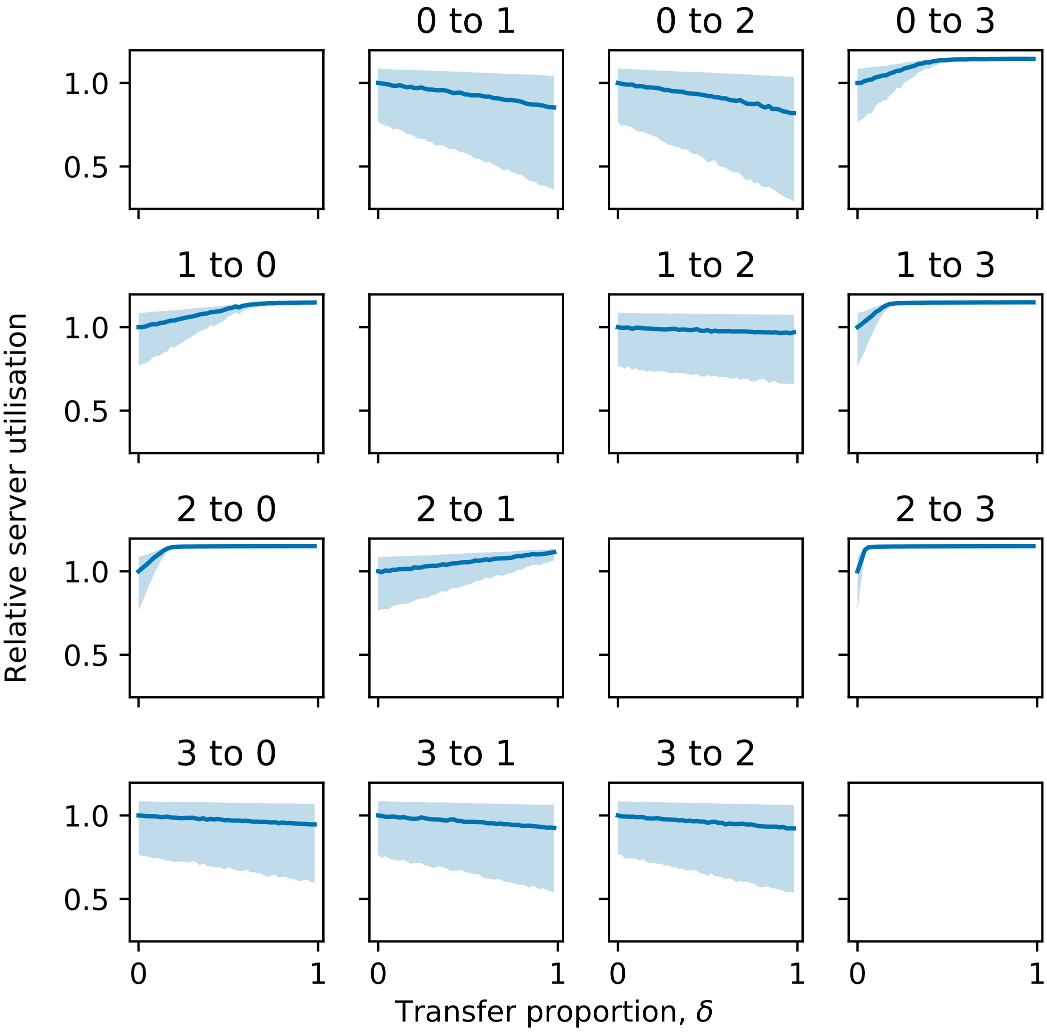


Figure 15. Plots of proportions of each cluster moving to another on relative server utilisation

	δ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Origin	**Destination**										
0	**1**	0.0	-0.0176	-0.0299	-0.0391	-0.0535	-0.0693	-0.0824	-0.1001	-0.1129	-0.1325
	2	0.0	-0.0197	-0.0290	-0.0488	-0.0627	-0.0782	-0.0919	-0.1140	-0.1384	-0.1592
1	**2**	0.0	-0.0035	-0.0108	-0.0108	-0.0180	-0.0181	-0.0249	-0.0256	-0.0302	-0.0357
3	**0**	0.0	-0.0060	-0.0132	-0.0137	-0.0206	-0.0274	-0.0320	-0.0380	-0.0422	-0.0494
	1	0.0	-0.0089	-0.0206	-0.0232	-0.0246	-0.0384	-0.0451	-0.0532	-0.0626	-0.0685
	2	0.0	-0.0100	-0.0184	-0.0254	-0.0314	-0.0443	-0.0542	-0.0504	-0.0649	-0.0714

Table 4. Proportional changes in median relative utilisation for selected cluster transfers

condition.

Conversely, the concern arises when either of the low resource requirement clusters moves to Cluster 0 or Cluster 3. Even as few as one in ten of the low-complexity, low-resource-needs arrivals in Cluster 2 moving to either cluster results in large jumps in the median system time for all arrivals. Soon after, as in the previous scenario, any variation in the system times disappears, indicating an overborne system.

With relative server utilisation, the story is much the same. The ordinary levels of high-complexity, high-resource arrivals from Cluster 3 are absorbed by the system and moving these arrivals to another cluster bears little effect on resource consumption levels. Likewise, either of the low-resource needs clusters moving even slightly toward high resource requirements completely overruns the system's resources. However, the relative utilisation levels of the system resources can be substantially reduced by moving arrivals from Cluster 0 to either Cluster 1 or Cluster 2, i.e. by reducing the overall resource requirements of such spells.

In essence, this entire analysis offers two messages. Firstly, that there are several ways in which the system can get worse and even overwhelmed. Secondly, and more importantly, that any meaningful impact on the system must come from a stimulus outside of the system that results in a higher proportion of healthy patients arriving at the hospital. This conclusion is non-trivial; the first two scenarios in this analysis show that there are no quick solutions to reduce the effect of COPD patients on hospital capacity and length of stay. The only effective intervention for improving the system on the whole is found through inter-cluster transfers.

5. Summary

This work presents a novel approach to investigating a healthcare population that encompasses the topics of segmentation analysis, queuing models, and the recovery of queuing parameters from incomplete data. This investigation is done despite characteristic limitations in operational research concerning the availability of fine-grained data, and the analysis in this manuscript only uses administrative hospital spell data from patients presenting COPD from Cwm Taf Morgannwg University Health Board.

By considering a variety of attributes present in the data, and engineering some, a useful clustering of the spell population is identified that successfully feeds into a multi-class $M/G/c$ queue to model a hypothetical COPD ward. This clustering was generated using the initialisation presented in Chapter **??**, which in turn was effectively evaluated using the EDO method from Chapter **??**. The culmination of these three features from this thesis fulfil its objective: to utilise machine learning through creation, evaluation and, finally, application.

With this model, several insights are gained by investigating purposeful changes in the parameters of the model that have the potential to inform actual public health policy. In particular, since neither the resource capacity of the system nor the clinical processes of the spells are evident in the data, service times and resource levels are not available. However, the length of stay is. Using what is available, this work assumes that mean service times can be parameterised using mean lengths of stay. By using the Wasserstein distance to compare the distribution of the simulated lengths of stay data with the observed data, a best performing parameter set is found via a parameter sweep.

This parameterisation ultimately recovers a surrogate for service times for each cluster, and a universal number of servers to emulate resource availability. The parameterisation itself offers its strengths by being straightforward and effective. Despite its simplicity, a good fit to the observed data is found, and — as is evident from the closing section of this manuscript — substantial and useful insights can be gained into the needs of the population under study.

This mode of analysis, in effect, considers all types of patient arrivals and how they each impact the system in terms of resource capacity and length of stay. By investigating changes in both overall patient arrivals and resource capacity, it is clear that there is no quick solution to be employed from within the hospital to improve COPD patient spells. The only effective, non-trivial intervention is to improve the overall health of the patients arriving at the hospital, as is shown by moving patient arrivals between clusters. In reality, this would correspond to an external, preventive policy that improves the overall health of COPD patients.

References

Arnolds, I. V., & Gartner, D. (2018). Improving hospital layout planning through clinical pathway mining. *Annals of Operations Research*, *263*, 453–477.

Asanjarani, A., Nazarathy, Y., & Pollett, P. (2017). *Parameter and state estimation in queues and related stochastic models: A bibliography.* Retrieved from `https://people.smp.uq.edu.au/PhilipPollett/papers/Qest/QEstAnnBib.pdf`

Bell, P. C., & O'Keefe, R. M. (1987). Visual interactive simulation — history, recent developments, and major issues. *SIMULATION*, *49*(3), 109–116.

Bhat, U. N. (2015). *An introduction to queueing theory.* Birkhäuser Boston.

Bhattacharjee, P., & Ray, P. K. (2014). Patient flow modelling and performance analysis of healthcare delivery processes in hospitals: A review and reflections. *Computers & Industrial Engineering*, *78*, 299–312.

Bittencourt, O., Verter, V., & Yalovsky, M. (2018). Hospital capacity management based on the queueing theory. *International journal of productivity and performance management.*, *67*(2), 224–238.

Brailsford, S. C., Bolt, T. B., Bucci, G., Chaussalet, T. M., Connell, N. A., Harper, P. R., ... Taylor, M. (2013). Overcoming the barriers: A qualitative study of simulation adoption in the NHS. *Journal of the Operational Research Society*, *64*(2), 157–168.

Cochran, J. K., & Roche, K. T. (2009). A multi-class queuing network analysis methodology for improving hospital emergency department performance. *Computers & Operations Research*, *36*(5), 1497–1512.

Collins, P. F., Stratton, R. J., Kurukulaaratchy, R. J., & Elia, M. (2018). Influence of deprivation on health care use, health care costs, and mortality in COPD. *International Journal of Chronic Obstructive Pulmonary Disease*, *13*, 1289–1296.

Dagkakis, G., & Heavey, C. (2016). A review of open source discrete event simulation software for operations research. *Journal of Simulation*, *10*(3), 193–206.

Delias, P., Doumpos, M., Grigoroudis, E., Manolitzas, P., & Matsatsinis, N. (2015). Supporting healthcare management decisions via robust clustering of event logs. *Knowledge-Based Systems*, *84*, 203–213.

Djabali, Y., Rabta, B., & Aissani, D. (2018). Approximating service-time distributions by phase-type distributions in single-server queues: A strong stability approach. *International Journal of Mathematics in Operational Research*, *12*, 507–531.

Erlang, A. K. (1917). Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Post Office Electrical Engineer's Journal*, *10*, 189–197.

Erlang, A. K. (1920). Telephone waiting times. *Matematisk Tidsskrift, B*, *31*, 25.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis.* John Wiley & Sons.

Fitzpatrick, B. G. (2019). Issues in reproducible simulation research. *Bulletin of Mathematical Biology*, *81*, 1–6.

Goldenshluger, A. (2016). Nonparametric estimation of the service time distribution in the M/G/∞ queue. *Advances in Applied Probability*, *48*(4), 1117–1138.

Hagenaars, J. A. (2002). *Applied latent class analysis.* Cambridge University Press.

Harper, P. R., & Winslett, D. (2006). Classification trees: A possible method for maternity risk grouping. *European Journal of Operational Research*, *169*, 146–156.

Houben-Wilke, S., Triest, F. J. J., Franssen, F. M., Janssen, D. J., Wouters, E. F., & Vanfleteren, L. E. (2019). Revealing methodological challenges in chronic obstructive pulmonary disease studies assessing comorbidities: A narrative review. *Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation*, *6*(2), 166–177.

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. In *The first Pacific-Asia conference on knowledge discovery and data mining* (pp. 21–34).

Ivie, P., & Thain, D. (2018). Reproducibility in scientific computing. *ACM Computing Surveys*, *51*(3).

Kingman, J. F. C. (2009). The first Erlang century—and the next. *Queueing Systems*, *63*(1-4), 3–12.

Komashie, A., Mousavi, A., Clarkson, P. J., & Young, T. (2015). An integrated model of patient and staff satisfaction using queuing theory. *IEEE Journal of Translational Engineering in Health and Medicine*, *3*, 1–10.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*(1), 79–86.

Kuwornu, J. P., Lix, L. M., & Shooshtari, S. (2014). Multimorbidity disease clusters in Aboriginal and non-Aboriginal Caucasian populations in Canada. *Chronic Diseases and Injuries in Canada*, *34*(4), 218–225.

Larsen, F. B., Pedersen, M. H., Friis, K., Glümer, C., & Lasgaard, M. (2017). A latent class analysis of multimorbidity and the relationship to socio-demographic factors and health-related quality of life. a national population-based study of 162,283 Danish adults. *PLoS One*, *12*(1).

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Houghton Mifflin Co.

McClain, J. O. (1976). Bed planning using queuing theory models of hospital occupancy: A sensitivity analysis. *Inquiry*, *13*(2), 167–176.

Mohammadi, A., & Salehi-Rad, M. R. (2012). Bayesian inference and prediction in an $M/G/1$ with optional second service. *Communications in Statistics - Simulation and Computation*, *41*(3), 419–435.

NHS Data Model and Dictionary. (n.d.-a). *NHS Business Definitions: Consultant Episode (Hospital Provider).* Retrieved from `https://www.datadictionary.nhs.uk/`

NHS Data Model and Dictionary. (n.d.-b). *NHS Business Definitions: Hospital Provider Spell.* Retrieved from `https://www.datadictionary.nhs.uk/`

Olafsson, S., Li, X., & Wu, S. (2008). Operations research and data mining. *European Journal of Operational Research*, *87*(3), 1429–1448.

Palmer, G. I., Knight, V. A., Harper, P. R., & Hawa, A. L. (2019). Ciw: An open-source discrete event simulation library. *Journal of Simulation*, *13*(1), 68–82.

Palvannan, R. K., & Teow, K. L. (2012). Queueing for healthcare. *Journal of Medical Systems*, *36*, 541–547.

Pinto, L. R., de Campos, F. C. C., Perpétuo, I. H. O., & Ribeiro, Y. C. N. M. B. (2014). Analysis of hospital bed capacity via queuing theory and simulation. In *Proceedings of the winter simulation conference 2014* (p. 1281-1292).

Ramdas, A., Trillos, N. G., & Cuturi, M. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, *19*(2), 47.

Rebuge, Á., & Ferreira, D. R. (2012). Business process analysis in healthcare environments: A methodology based on process mining. *Information Systems*, *37*(2), 99–116.

Sexton, E., & Bedford, D. (2016). GP supply, deprivation and emergency admission to hospital for COPD and diabetes complications in counties across Ireland: An exploratory analysis.

Irish Journal of Medical Science, *185*(2), 453–461.

Shortle, J. F., Thompson, J. M., Gross, D., & Harris, C. M. (2018). *Fundamentals of queueing theory.* John Wiley & Sons, Inc.

Simon-Tuval, T., Scharf, S. M., Maimon, N., Bernhard-Scharf, B. J., Reuveni, H., & Tarasiuk, A. (2011). Determinants of elevated healthcare utilization in patients with COPD. *Respiratory Research*, *12*(7).

Steiner, M. C., Lowe, D., Beckford, K., Blakey, J., Bolton, C. E., Elkin, S., ... Singh, S. J. (2017). Socioeconomic deprivation and the outcome of pulmonary rehabilitation in England and Wales. *Thorax*, *72*(6), 530–537.

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

Stewart, W. J. (2009). *Probability, markov chains, queues, and simulation.* Princeton University Press.

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

Vuik, S. I., Mayer, E. K., & Darzi, A. (2016a). Patient segmentation analysis offers significant benefits for integrated care and support. *Health Affairs*, *35*(5), 769–775.

Vuik, S. I., Mayer, E. K., & Darzi, A. (2016b). A quantitative evidence base for population health: Applying utilization-based cluster analysis to segment a patient population. *Population Health Metrics*, *14*.

Williams, J., Dumont, S., Parry-Jones, J., Komenda, I., Griffiths, J., & Knight, V. (2015). 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.

Wu, X., & Kumar, V. (2009). *The top ten algorithms in data mining.* CRC press.

Yan, S., Kwan, Y. H., Tan, C. S., Thumboo, J., & Low, L. L. (2018). A systematic review of the clinical application of data-driven population segmentation analysis. *BMC Medical Research Methodology*, *18*(121).

Yan, S., Seng, B. J. J., Kwan, Y. H., Tan, C. S., Quah, J. H. M., Thumboo, J., & Low, L. L. (2019). Identifying heterogeneous health profiles of primary care utilizers and their differential healthcare utilization and mortality – a retrospective cohort study. *BMC Family Practice*, *20*(54).

Yom-Tov, G. B., & Mandelbaum, A. (2014). Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, *16*(2), 283–299.

Yoon, S., Goh, H., Kwan, Y. H., Thumboo, J., & Low, L. L. (2020). Identifying optimal indicators and purposes of population segmentation through engagement of key stakeholders: A qualitative study. *Health Res Policy Syst.*, *18*(1), 26.