IMA Journal of Management Mathematics (2025) **36**, 255–274
https://doi.org/10.1093/imaman/dpaf002
Advance Access publication on 9 January 2025

Optimizing healthcare queues: a case study on chronic respiratory illness

JONATHAN GILLARD[†] AND VINCENT KNIGHT
School of Mathematics, Cardiff University, Cardiff CF10 3AX, UK
[†]Corresponding author. Email: gillardjw@cardiff.ac.uk

KENDAL SMITH
Cwm Taf Morgannwg University Health Board, National Health Service, London E1 8EU, UK

AND

HENRY WILDE
Data Science Campus, Office for National Statistics, Newport NP10 8XG, UK

[Received on 10 January 2024; accepted on 13 December 2024]

Accepted by: M. Zied Babai

This study employs a data-driven approach to assess the evolving resource needs of chronic obstructive pulmonary disease (COPD) patients, exploring the impact on the hospital system. It integrates segmentation, operational queuing theory and parameter recovery from incomplete data to overcome limitations in fine-grained data availability, yielding operational insights using only administrative data. Initiating with a population clustering from granular data, the paper utilizes a multi-class $M/M/c$ model, extracting parameters through parameterization and Wasserstein distance. This model facilitates an informative analysis of the queuing system and population needs through various what-if scenarios. The comprehensive analyses encompass all patient arrival types, revealing that addressing the impact of COPD patients on the system necessitates more than just expanding capacity. Our work demonstrates the potential for specific improvement in clinical performance in respect of COPD patients.

Keywords: OR in health services; machine learning; queueing.

1. Introduction

Population health research is increasingly based on data-driven methods (as opposed to those designed solely by clinical experts) for patient-centred care through 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. This work demonstrates a method of overcoming this, using routinely gathered, administrative hospital data to build a clustering that feeds into a multi-class queuing model, allowing for better understanding of the healthcare population and the system with which they interact.

Specifically, this work examines records of patient spells from the National Health Service (NHS) Wales Cwm Taf Morgannwg University Health Board (UHB) presenting chronic obstructive pulmonary disease (COPD). COPD is a condition of particular interest to population health research (Demir *et al.* (2009)), and to Cwm Taf Morgannwg UHB, 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 the Cwm Taf Morgannwg UHB had the highest prevalence of the condition across all the Welsh health boards. Operationally, the management of

COPD requires substantial resources, including frequent hospital admissions, long-term medication and specialist care, which strain hospital capacities and increase healthcare costs. Socially, COPD patients often experience reduced quality of life and mobility, necessitating comprehensive social support services to assist with daily activities and mental health care, further burdening community healthcare services. Medically, COPD is associated with high morbidity and comorbidities, such as heart disease and diabetes, complicating treatment protocols and requiring multidisciplinary approaches. These combined factors underscore the need for robust, integrated healthcare strategies to effectively manage the widespread impact of COPD.

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 the paper is structured as follows: Section 1 provides a literature review, and an overview of the dataset and its clustering; Section 2 describes the queuing model used and the estimation of its parameters; Section 3 presents several what-if scenarios with insight provided by the model parameterization and the clustering; Section 4 offers some managerial insight into our results of our work before Section 5 concludes the paper. We seek to describe the managerial implications of our work throughout.

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 (Benton & Hand (2002)) 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 *et al.* (2016a); Yoon *et al.* (2020)). For instance, clinical profiling often forms part of the broader analysis where each segment is summarized in a phrase or infographic (Vuik *et al.* (2016b); Yan *et al.* (2019)).

The review for this work identified three commonplace groups of patient characteristics used to segment a patient population: system utilization 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 & Gartner (2018) and Delias *et al.* (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 *et al.* (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 (Lazarsfeld & Henry (1968); Hagenaars (2002)). This method has proved useful in the study of comorbidity patterns

as in Kuwornu *et al.* (2014); Larsen *et al.* (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 maximize homogeneity within, and heterogeneity between, each cluster (Everitt *et al.* (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 *et al.* (2008); Wu & 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 & 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 (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 Crowe *et al.* (2012); Palvannan & Teow (2012); Pinto *et al.* (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) devise new categorizations for both hospital beds and arrivals that are informed by the queuing model. A further example is Komashie *et al.* (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 (Rebuge & Ferreira (2012); Arnolds & Gartner (2018)), and patient flow (Bhattacharjee & 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 organizations 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 & 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 *et al.* (2019)). Ciw is a discrete event simulation library written in Python that is fully documented and tested. The simulations constructed and studied in Sections 2 and 3 utilize 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 are 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 *et al.* (2021) 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 & Salehi-Rad (2012) and Yom-Tov & 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 & 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 & Mandelbaum (2014) consider 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. Goldenshluger (2016) and Djabali *et al.* (2018) are two such recent works that consider an underlying process to estimate a general service time distribution in single server and infinite server queues respectively.

1.1.4. *Critical analysis of current literature.* The techniques discussed, while valuable, have limitations that must be critically examined. Segmentation analysis, while helpful in identifying distinct patient groups, risks oversimplification. Unsupervised methods like LCA and clustering rely heavily on the quality of available data and predefined assumptions, which can lead to biased or incomplete results. Moreover, the focus on operational efficiency, as highlighted in the use of process mining for system optimization (Delias *et al.* (2015); Arnolds & Gartner (2018)), can sideline the more pressing issue of patient-centred care.

In queuing models, the historical focus on service capacity (McClain (1976)) and the development of actionable outputs (Pinto *et al.* (2014)) have yielded significant improvements in hospital operations. However, the limitations of theoretical models and the challenges posed by incomplete data, as seen in healthcare-specific research such as Mohammadi & Salehi-Rad (2012); Yom-Tov & Mandelbaum (2014), restrict the practical impact of such approaches. The introduction of computer simulation models addresses some of these gaps but introduces new challenges around reproducibility and the need for robust, open-source tools see Fitzpatrick (2019) and Palmer *et al.* (2019).

Finally, the handling of incomplete queue data highlights a major challenge in healthcare research. The scarcity of data often results in the termination of studies before meaningful insights can be drawn (Asanjarani *et al.* (2021)). This issue not only limits the development of accurate queuing models but also affects decision-making processes in resource allocation and service planning. Future research must address these gaps by prioritizing the collection of high-quality data and promoting the use of open-source tools that enhance transparency and reproducibility in modelling and simulation.

1.2. *Overview of the dataset and its clustering*

The Cwm Taf Morgannwg UHB provided the dataset used in this work. The dataset contains an administrative summary of 5,231 patients presenting COPD from February 2011 through March 2019

totalling 10,861 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. The following attributes describe the spells included in the dataset:

- Personal identifiers and information, i.e. patient and spell ID numbers, and identified gender;
- Admission/discharge dates and approximate times;
- Attributes summarizing 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 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 conditions (LTCs) as well as indicators for some other conditions such as sepsis and obesity. CCI is useful in anticipating hospital utilization 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 (Sexton & Bedford (2016); Steiner *et al.* (2017); Collins *et al.* (2018)).

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 another administrative dataset provided by the Cwm Taf Morgannwg UHB;
- The presenting ICD codes were generalized 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 12 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. The clustering algorithm of choice is a variant of k-means, called k-prototypes, allows for the clustering of mixed-type data by performing k-means on the numeric attributes and k-modes on the categoric.

The attributes included in the clustering encompass both utilization metrics and clinical attributes relating to the spell. They comprise the summative 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).

To determine the optimal number of clusters, k, the knee point detection algorithm introduced in Satopaa *et al.* (2011) was used with a range of potential values for k from two to 10. This range was chosen based on what may be considered feasibly informative to stakeholders. The knee point detection algorithm can be considered a deterministic version of the widely known ‘elbow method’ for determining the number of clusters. Applying this algorithm revealed an optimal value for k of four. The initialization method used for k-prototypes was presented in Gillard *et al.* (2023) as it was found to give an improvement in the clustering over other initialization methods.

A summary of the spells is provided in Table 1. 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 summarized 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 7 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 1–5 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. As may be expected, broadly, each cluster refers to the severity of the condition of the clustered patients.

Figure 1 shows the length of stay distributions as histograms. Figure 1(a) demonstrates the different bed resource requirements well for each cluster—better than Table 1 might — in that the difference between the clusters is not just 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 Fig. 1(b), 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 visualization due to the widespread distribution of their lengths of stay.

The same conclusions can be drawn about spell costs from Fig. 2; there are distinct patterns between the clusters in terms of their costs, and they align with the patterns seen in Fig. 1. 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, Fig. 3 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

TABLE 1 *A summary of clinical and condition-specific characteristics for each cluster and the population. A negative length of stay indicates that the patient died prior to arriving at the hospital.*

		Cluster				Population
		0	1	2	3	
Characteristics	**Percentage of spells**	9.91	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	-0.00	-0.02	48.82	-0.02
	Mean LOS	25.30	6.46	4.11	75.36	7.68
	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
	Inter-arrival rate	2.07	3.22	3.35	2.30	3.14
Intervention prevalence	**None, %**	80.20	83.42	65.76	89.74	70.94
	PR, %	15.80	13.43	27.97	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, %	19.05	28.14	14.84	25.00	17.96
	AMI, %	13.85	22.93	8.76	16.03	12.10
	CHF, %	12.45	53.85	0.00	26.28	11.99
	Renal disease, %	7.53	19.54	1.92	17.95	6.10
	Cancer, %	7.62	12.23	2.93	10.90	5.30
	Dementia, %	6.88	21.26	0.00	26.92	5.17
	CVA, %	8.64	13.33	0.70	19.87	4.20
	PVD, %	4.37	7.69	2.27	5.77	3.57
	CTD, %	5.11	4.25	3.11	4.49	3.54
	Obesity, %	2.51	3.01	1.49	7.69	1.97
	Metastatic cancer, %	1.58	4.49	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

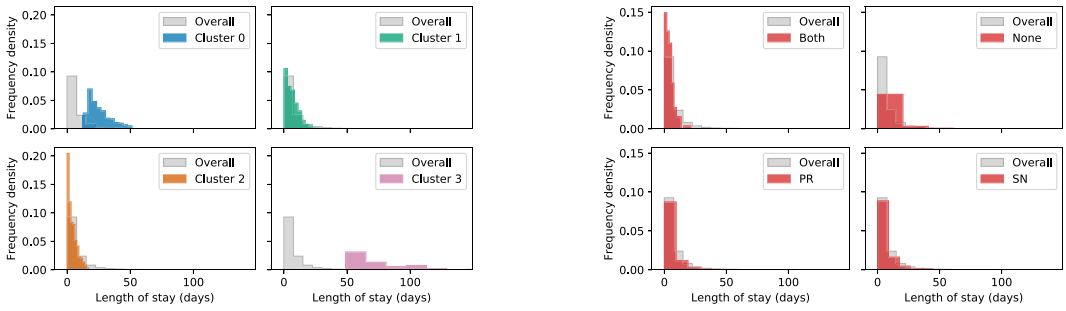


FIG. 1. Histograms for length of stay by (a) cluster and (b) intervention.

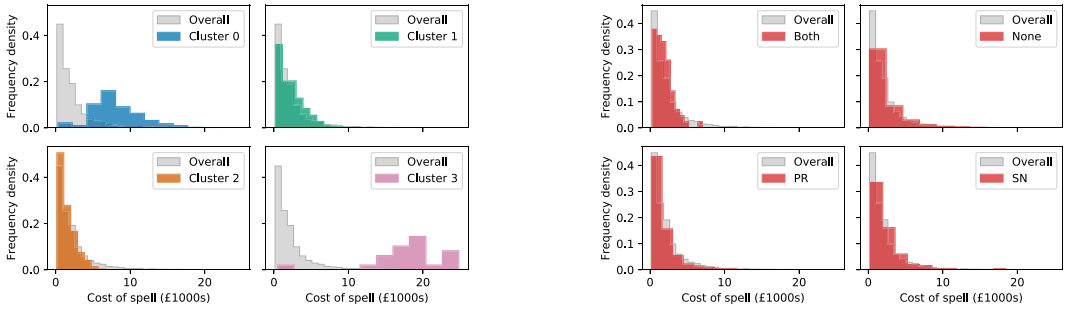


FIG. 2. Histograms for spell cost by (a) cluster and (b) intervention.

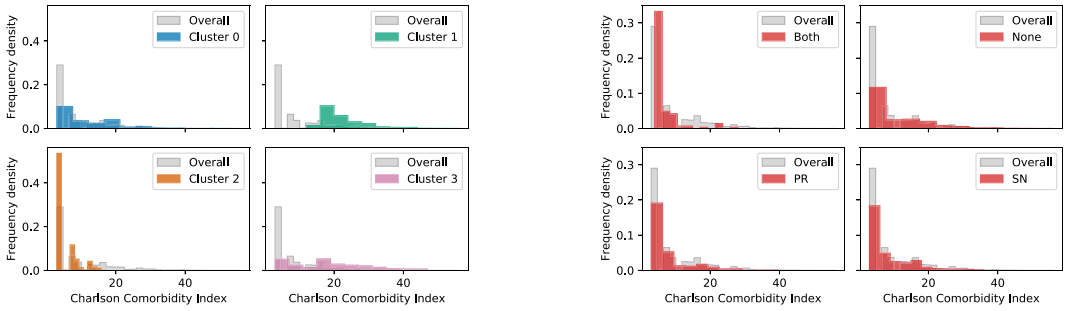


FIG. 3. Histograms for CCI by (a) cluster and (b) intervention.

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 Fig. 3(b) all display similar, highly skewed distributions regardless of intervention.

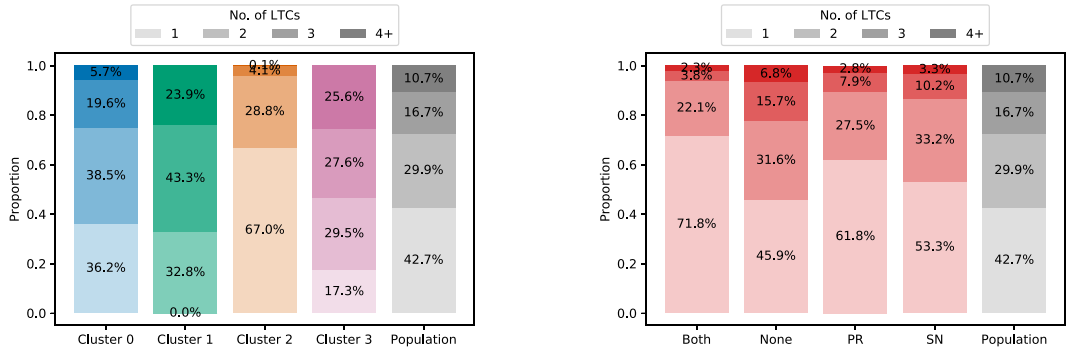


FIG. 4. Proportions of the number of concurrent LTCs in a spell by (a) cluster and (b) intervention.

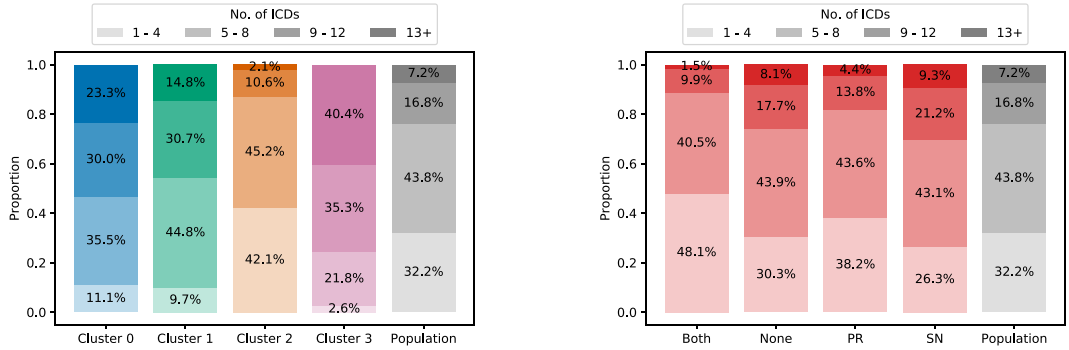


FIG. 5. Proportions of the number of concurrent ICDs in a spell by (a) cluster and (b) intervention.

Figures 4 and 5 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 Fig. 4(a), 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 severely biased towards patients presenting COPD as the only LTC; Cluster 3 is the most uniformly spread among the four bins despite the increased length of stay and CCI suggesting a diverse array of patients in terms of their long term medical needs.

Figure 5(a) largely mirrors these cluster profiles with the number of concurrent ICDs. Some points of interest, however, are that Cluster 1 has a relatively low-leaning distribution of ICDs that does not marry up with the high rates of LTCs, and that 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. Fig. 4(b) and 5(b)), 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 work 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.

2. Constructing the queuing model

The scarcity of data limits the options for the queuing model. However, there is a precedent for simplifying healthcare systems to a single node with parallel servers that emulate resource availability. Steins & Walther (2013) and 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, 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.

In order to follow in the suit of recent literature, 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 despite the system time distributions shown in Fig. 1(a) in order to simplify the model parameterization. Also note the work of Takacs (1969), which states that $M/G/c$ queues are well-approximated by $M/M/c$ queues.
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.
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 reciprocal of the mean inter-arrival times for that group. This parameter is denoted by λ_i for each cluster i. 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. For simplicity, this work assumes that for each cluster, i, the mean service time of that cluster, $\frac{1}{\mu_i}$, is directly proportional to the mean total system time of that cluster, $\frac{1}{\phi_i}$, such that:

$$\mu_i = \phi_i / p_i \tag{1}$$

where $p_i \in]0, 1]$ is some parameter to be determined for each group.

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 & 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 just from 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 the typical Euclidean distance), and its values take the units of the distributions under comparison (in this case: days). 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 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:

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

where F and G are the cumulative density functions of U and V, respectively. Statement of (2) is presented in Ramdas *et al.* (2017). The parameter set with the smallest maximum distance between the simulated system time distribution and the overall observed length of stay distribution is then taken to be the most appropriate.

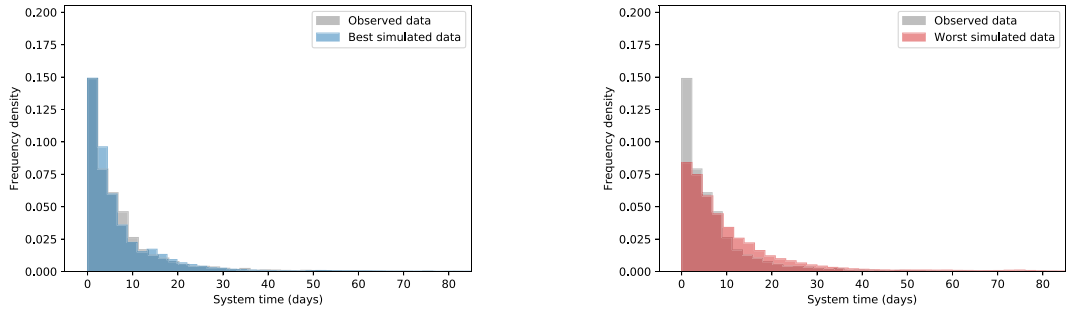
We compute the worst-case Wasserstein distance over a number of simulations, that is, we seek $\max_{s\in S} W(T_{c,p}, T)$ where S is a set of random simulation seeds and W is the Wasserstein distance (2), where T denotes the system time distribution of all of the observed data and $T_{c,p}$ denotes the system time distribution obtained from a simulation with c servers and $p := \left(p_0, p_1, p_2, p_3\right)$. Then the optimal parameter set (c^*, p^*) is given by

$$\left(c^*, p^*\right) = \arg\min_{c,p} \left\{ \max_{s\in S} W(T_{c,p}, T) \right\} \tag{3}$$

Each trial 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. Any metric of the simulated Wasserstein distances could be evaluated in (3), but choosing the worst-case scenario encourages a robust estimation of queueing parameters c and p. Other metrics, such as median, would not afford us this property. For example, in our analyses, use of the median inflated the presence of short-term patients. Each parameter set was repeated 50 times, with each simulation running for 4 years of virtual time. The warm-up and cool-down periods were taken to be approximately 1 year each, leaving two years of simulated data from each repetition. Justification for this approach is that 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, our approach has been to simulate the length of stay distribution for a range of values p_i and c, to find the parameters that best match the observed data.

TABLE 2 *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).*

	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	7.68	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



(a) (b)

FIG. 6. Histograms of the simulated and observed length of stay data for the (a) best and (b) worst parameter sets.

The parameter sweep included values of each p_i from 0.5 to 1.0 with a granularity of 5.0×10^{-2} and values of c from 40 to 60 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. The range and granularity for the search for c was informed by practitioner advice. Each parameter set was repeated 50 times with each simulation running for 4 years of virtual time. The warm-up and cool-down periods were taken to be approximately 1 year each leaving two years of simulated data from each repetition.

The results of this parameter sweep can be summarized in Fig. 6. Each plot 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. In the best case, a very close fit has been found. Meanwhile, Fig. 6(b) 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 2 reinforces these results numerically, showing a precise fit by the best parameters across the board. Note that, as stated earlier, p is constrained to have an upper bound of 1. This is due to the fact that the overall length of stay must be longer than the length of service. Allowing the parameter p to be larger than 1 would lose the viability of the $M/M/c$ model.

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.

3. Adjusting the queuing model

This section comprises several what-if scenarios—a classic component of healthcare operational research—under the novel parameterization of the queue established in Section 2. The outcomes of interest in this work are server (resource) utilization 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 would the system be affected by a change in overall patient arrivals?
- How is the system affected by a change in resource availability (i.e. a change in c)?
- How is the system affected by patients moving between clusters?

Given the nature of the observed data, the queuing model parameterization and its 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'.

3.1. *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 $\sigma \in \mathbb{R}$. This scaling factor is applied to the model by multiplying each cluster's arrival rate by σ. That is, for cluster i, its new arrival rate, $\hat{\lambda}_i$, is given by

$$\hat{\lambda}_i = \sigma \lambda_i \tag{4}$$

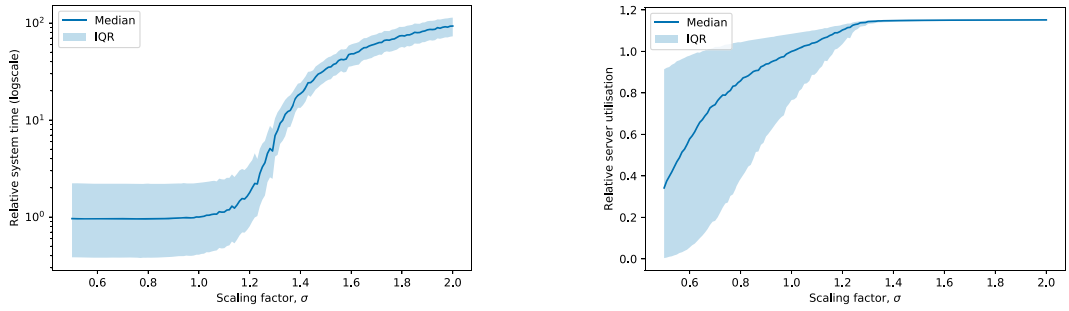
Figure 7 shows the effects of changing patient arrivals on (a) relative system times and (b) relative server utilization for values of σ from 0.5 to 2.0 at a precision of 1.0×10^{-2}. Specifically, each plot in the figure (and the subsequent figures in this section) shows the median and interquartile range (IQR) of each relative attribute. These metrics provide an insight into the experience of the average user (or server) in the system. Furthermore, they reveal the stability or variation of the body of users (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 Fig. 7(a), for instance, the relative system times (i.e. the relative length of stay for patients) remains unchanged for a range of small σ.

However, Fig. 7(b) 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 $\sigma \approx 1.4$). That is, the median server utilization reaches a maximum (this corresponds to constant utilization) at this point, and the variation in server utilization disappears entirely.

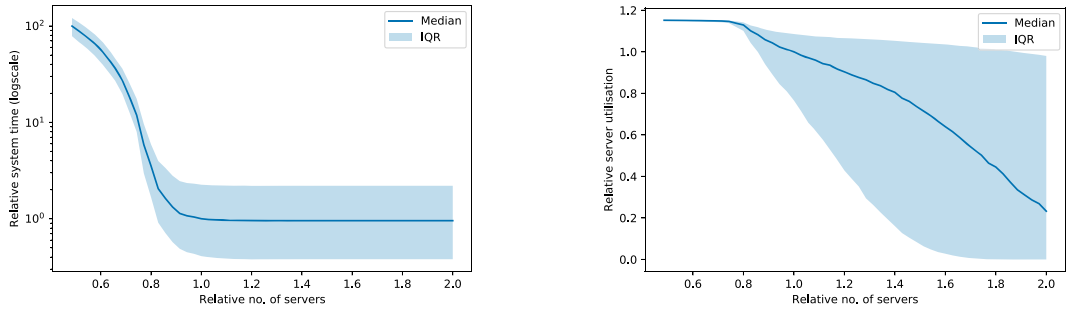
3.2. *Changes to resource availability*

As is discussed in Section 2, 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



(a) (b)

FIG. 7. Plots of σ against relative (a) system time and (b) server utilization.



(a) (b)

FIG. 8. Plots of the relative number of servers against relative (a) system time and (b) server utilization.

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 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 easily seen, and do away with any concerns as to what a particular number of servers precisely reflects in the real world.

Figure 8 shows how the relative resource availability affects relative system times and server utilization. In this scenario, the relative number of servers took values from 0.5 to 2.0 at steps of 2.5×10^{-2} — this is equivalent to a step size of one in the actual number of servers. 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.

Moreover, the variation in the body of the relative times (i.e. the IQR) decreases as resource availability decreases. 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.

3.3. *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 characterized 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 enact that movement informed by a model such as this, then real change would be seen in the real system. See for example Saha & Ray (2019) for further details on this issue.

In order to model the effects of spells moving between two clusters, the assumption is that services remain the same (and so does each cluster's p_i), but their arrival rates are altered according to some transfer proportion. Consider two clusters indexed at i, j, and their respective arrival rates, λ_i, λ_j, and let $\delta \in [0, 1]$ denote the proportion of arrivals to be moved from cluster i to cluster j. Then the new arrival rates for each cluster, denoted by $\hat{\lambda}_i, \hat{\lambda}_j$ respectively, are

$$\hat{\lambda}_i = (1 - \delta)\,\lambda_i \quad \text{and} \quad \hat{\lambda}_j = \delta\lambda_i + \lambda_j \tag{5}$$

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 utilization and system time can be measured independently.

Figures 9 and 10 show the effect of moving patient arrivals between clusters on relative system time and relative server utilization, respectively. 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 1.0 at steps of 2.0×10^{-2}.

Considering Fig. 9, it is clear that there are some cases where reducing particular types of spells (by making them like another type of spell) does not affect overall system times. Namely, moving the high resource requirement spells that describe Cluster 0 and Cluster 3 to any other cluster. These clusters make up only 10% of all arrivals, and this figure shows that in terms of system times, the model can handle them without concern under normal conditions. The concern comes when either of the other clusters moves to Cluster 0 or Cluster 3. Even as few as one in five 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, and soon after, as, in the previous scenario, any variation in the system times disappears indicating an overborne system.

With relative server utilization, 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 no 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 utilization levels of the system resources can be 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: that there are several ways in which the system can get worse and even overwhelmed but, more importantly, that any meaningful impact on the system must come from a stimulus outside of the system that results in healthier 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 or length of stay. The only effective intervention is found through inter-cluster transfers.

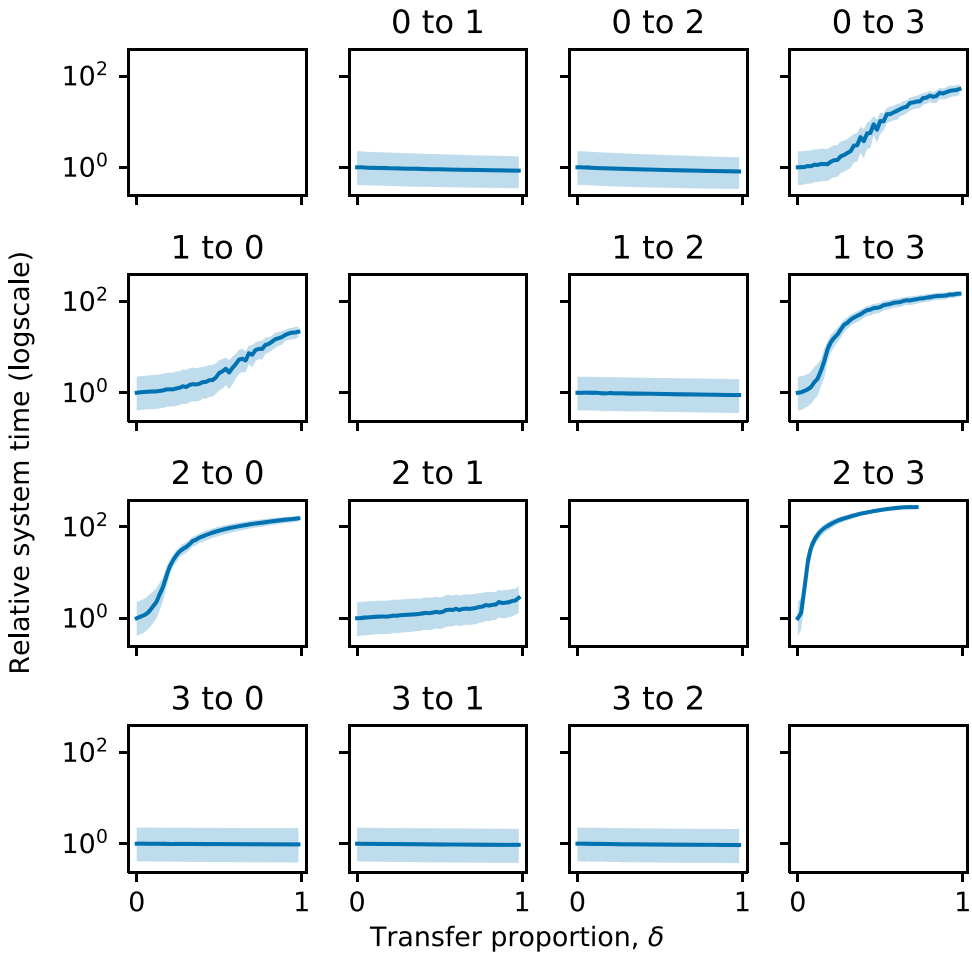


FIG. 9. Plots of proportions of each cluster moving to another against relative system time.

4. Managerial implications

The insights gained from the clustering of COPD patients and the application of queuing theory offer significant managerial implications for healthcare service delivery. Effective management of healthcare systems, particularly in dealing with chronic conditions like COPD, requires an integrated and data-driven approach. This study highlights several critical areas where managerial decisions can lead to improved outcomes in both resource allocation and patient care.

4.1. *Resource allocation and capacity management*

The results demonstrate that simply adding service capacity, such as increasing the number of available beds or staff, does not necessarily lead to better patient outcomes in terms of system time. Managers should focus not only on increasing resources but also on optimizing the allocation of existing resources.

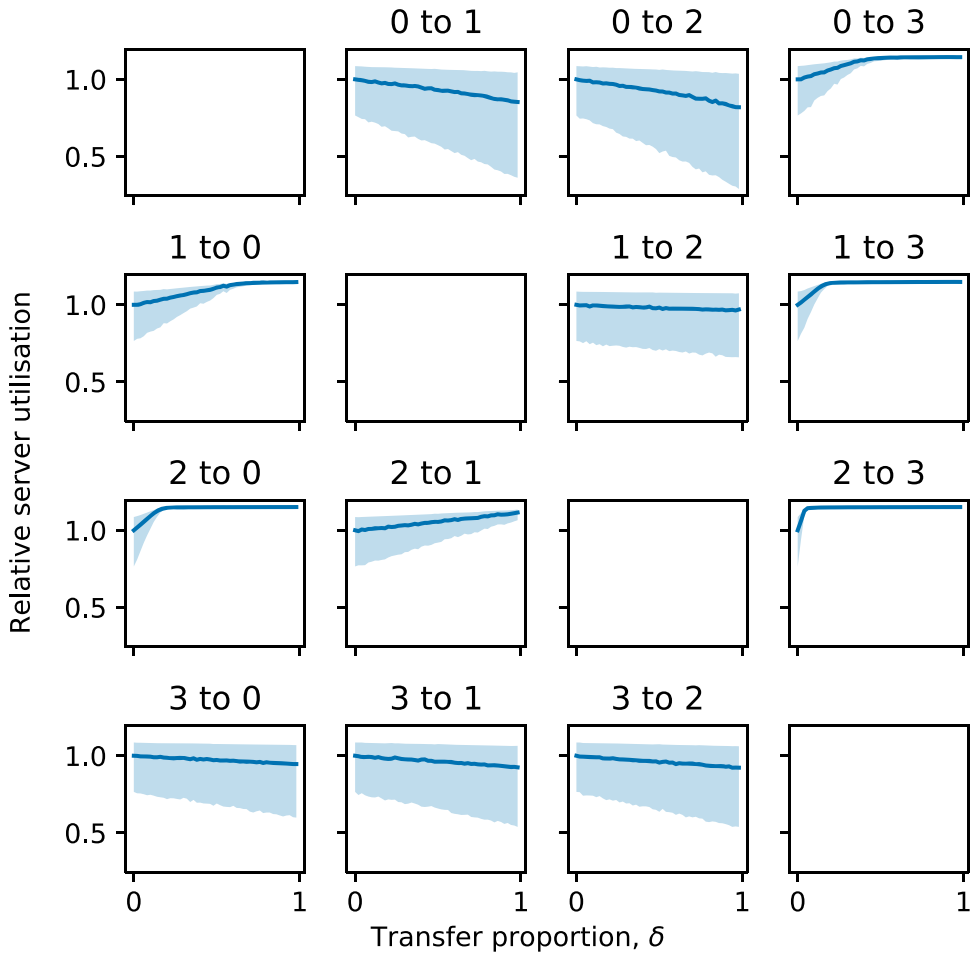


FIG. 10. Plots of proportions of each cluster moving to another on relative server utilization.

By redistributing patient load across clusters with varying levels of complexity, healthcare administrators can ensure that resources are used efficiently, reducing the risk of overburdening certain areas of care while underutilizing others.

4.2. *Cluster-specific resource planning*

Each identified patient cluster has different resource needs, clinical complexities and system times. Managers should tailor healthcare services to these distinct needs by allocating personnel, equipment and medications according to the specific characteristics of each cluster. For example, clusters with high comorbidity burdens, such as Cluster 3, require more intensive care and specialist resources. In contrast, patients in lower-resource clusters may benefit more from community-based or outpatient interventions.

4.3. *Preventive public health strategies*

The study's findings suggest that external interventions to improve patient health before hospital admission are critical for alleviating the system's overall burden. For instance, public health initiatives targeting early COPD management, lifestyle changes and community care programmes could significantly reduce hospital admissions and length of stay for high-risk patients. Hospital managers can collaborate with public health officials to integrate such strategies, thereby reducing the strain on acute care services.

4.4. *Data-driven decision making*

The integration of clustering techniques and queuing models offers managers a powerful tool for making evidence-based decisions. By using administrative data to model patient flows and service times, healthcare organizations can identify potential bottlenecks, predict patient demand and plan accordingly. This approach also provides flexibility in adjusting policies based on real-time data, enabling more responsive and adaptive healthcare management.

4.5. *Verification and validation of models*

To ensure the models' effectiveness, healthcare managers should implement verification and validation processes by testing the models on smaller samples of patient clusters. This will not only enhance the model's reliability but also provide a framework for continual improvement in resource planning and policy implementation. These processes are crucial in translating theoretical models into practical, actionable strategies that align with real-world challenges.

5. Conclusions

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 this work only uses administrative hospital spell data from patients presenting COPD from the Cwm Taf Morgannwg UHB. Further study is necessary to ascertain the impact of the Markovian (exponential) distribution for length of stay, since we approximate a likely $M/G/c$ queue with a $M/M/c$ one. Here a necessary choice has been made to balance parameterization complexity with data fit but further work will investigate the choice of more parameter heavy distributions.

Acknowledgements

The authors wish to thank the Cwm Taf Morgannwg University Health Board for their funding and support of the PhD of which this work has formed a part. We thank the two anonymous reviewers for their careful reading of the paper and their valuable comments.

Data availability statement

The data underlying this article cannot be shared publicly for the privacy of the individuals. A synthetic analogue has been archived and is available at https://github.com/daffidwilde/copd-paper along with all the source code used.

References

Arnolds, I. V. & Gartner, D. (2018) Improving hospital layout planning through clinical pathway mining. *Ann. Oper. Res.*, **263**, 453–477.

Asanjarani, A., Nazarathy, Y. & Taylor, P. (2021) A survey of parameter and state estimation in queues. *Queueing Syst.*, **97**, 39–80.

Benton, T. C. and Hand, D. J. (2002). Segmentation into predictable classes. *IMA J. Manag. Math.*, **13**(4): 245–259.

Bhattacharjee, P. & Ray, P. K. (2014) Patient flow modelling and performance analysis of healthcare delivery processes in hospitals: a review and reflections. *Comput. Ind. Eng.*, **78**, 299–312.

Brailsford, S. C., Bolt, T. B., Bucci, G., Chaussalet, T. M., Connell, N. A., Harper, P. R., Klein, J. H., Pitt, M., and Taylor, M. (2013). Overcoming the barriers: a qualitative study of simulation adoption in the NHS. *J. Oper. Res. Soc.*, **64**(2): 157–168.

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. *Int. J. Chron. Obstruct. Pulmon. Dis.*, **Volume 13**, 1289–1296.

Crowe, S., Pagel, C., Bull, K., Fenton, M., Vasilakis, C., Gallivan, S., and Utley, M. (2012). Using a mathematical model to assist with the management of paediatric heart transplant waiting lists: a case study. *IMA J. Manag. Math.*, **23**(2): 99–116.

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

Demir, E., Chaussalet, T., Xie, H., and Millard, P. H. (2009). Modelling risk of readmission with phase-type distribution and transition models. *IMA J. Manag. Math.*, **20**(4): 357–367.

Djabali, Y., Rabta, B. & Aissani, D. (2018) Approximating service-time distributions by phase-type distributions in single-server queues: a strong stability approach. *Int. J. Math. Oper. Res.*, **12**, 507–531.

Erlang, A. K. (1917) Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Post Office Electr. Eng. J.*, **10**, 189–197.

Erlang, A. K. (1920) Telephone waiting times. *Mat. Tidsskr. B*, **31**, 25.

Everitt, B. S., Landau, S., Leese, M. & Stahl, D. (2011) *Cluster Analysis*. New York: Wiley.

Fitzpatrick, B. G. (2019) Issues in reproducible simulation research. *Bull. Math. Biol.*, **81**, 1–6.

Gillard, J., Knight, V., and Wilde, H. (2023). A novel initialisation based on hospital-resident assignment for the *k*-modes algorithm. *Soft Comput.*, **27**(14): 9441–9457.

Goldenshluger, A. (2016). Nonparametric estimation of the service time distribution in the M/G/∞ queue. *Adv. Appl. Probab.*, **48**(4): 1117–1138.

Hagenaars, J. A. (2002) *Applied Latent Class Analysis*. New York: Cambridge University Press.

Harper, P. R. & Winslett, D. (2006) Classification trees: a possible method for maternity risk grouping. *Eur. J. Oper. Res.*, **169**, 146–156.

Houben-Wilke, S., Triest, F. J. J., Franssen, F. M., Janssen, D. J., Wouters, E. F., and Vanfleteren, L. E. (2019). Revealing methodological challenges in chronic obstructive pulmonary disease studies assessing comorbidities: a narrative review. *Chron. Obstruct. Pulmon. Dis.*, **6**(2): 166–177.

Ivie, P. and Thain, D. (2018). Reproducibility in scientific computing. *ACM Comput. Surv.*, **51**(3).

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

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.*, **22**(1): 79–86.

Kuwornu, J. P., Lix, L. M., and Shooshtari, S. (2014). Multimorbidity disease clusters in Aboriginal and non-Aboriginal Caucasian populations in Canada. *Chron. Dis. Inj. Canada*, **34**(4): 218–225.

Larsen, F. B., Pedersen, M. H., Friis, K., Glümer, C., and 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), e0169426.

Lazarsfeld, P. F. & Henry, N. W. (1968) *Latent Structure Analysis*. Boston, Massachusetts: 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. and Salehi-Rad, M. R. (2012). Bayesian inference and prediction in an $M/G/1$ with optional second service. *Commun. Stat. - Simul. Comput.*, **41**(3): 419–435.

Olafsson, S., Li, X., and Wu, S. (2008). Operations research and data mining. *Eur. J. Oper. Res.*, **87**(3): 1429–1448.

Palmer, G. I., Knight, V. A., Harper, P. R., and Hawa, A. L. (2019). Ciw: an open-source discrete event simulation library. *J. Simul.*, **13**(1): 68–82.

Palvannan, R. K. & Teow, K. L. (2012) Queueing for healthcare. *J. Med. Syst.*, **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. *Proc. Winter Simul. Conf.*, **2014**, 1281–1292.

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

Rebuge, Á. and Ferreira, D. R. (2012). Business process analysis in healthcare environments: a methodology based on process mining. *Inf. Syst.*, **37**(2): 99–116.

Saha, E. & Ray, P. K. (2019) Modelling and analysis of inventory management systems in healthcare: a review and reflections. *Comput. Ind. Eng.*, **137**, 106051.

Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. (2011) Finding a 'kneedle' in a haystack: detecting knee points in system behavior. *Proceedings of the 2011 31st International Conference on Distributed Computing Systems Workshops*, pp. 166–171.

Sexton, E. and Bedford, D. (2016). GP supply, deprivation and emergency admission to hospital for COPD and diabetes complications in counties across Ireland: an exploratory analysis. *Ir. J. Med. Sci.*, **185**(2): 453–461.

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

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

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

Takacs, L. (1969). On Erlang's formula. *Ann. Math. Stat.*, **40**(1): 71–78.

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

Vuik, S. I., Mayer, E. K., and 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. *Popul. Health Metr.*, **14**.

Williams, J., Dumont, S., Parry-Jones, J., Komenda, I., Griffiths, J., and 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. and Kumar, V. (2009). *The Top Ten Algorithms in Data Mining*. Boca Raton, Florida: CRC Press, Chapman and Hall/CRC.

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

Yan, S., Seng, B. J. J., Kwan, Y. H., Tan, C. S., Quah, J. H. M., Thumboo, J., and 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 Pract.*, **20**(54), 54.

Yom-Tov, G. B. and Mandelbaum, A. (2014). Erlang-R: a time-varying queue with reentrant customers, in support of healthcare staffing. *Manuf. Serv. Oper. Manag.*, **16**(2): 283–299.

Yoon, S., Goh, H., Kwan, Y. H., Thumboo, J., and 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.