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 28, 2022

ABSTRACT
This work uses a data-driven approach to analyse how the requirements of patients with chronic obstructive pulmonary disease (COPD) may change, quantifying their impact on the hospital system with which they interact. This approach is composed of a novel combination of often distinct modes of analysis: segmentation, queuing theory, and parameter recovery. Through this combination of methods, this work demonstrates how to overcome potential limitations presented by a lack of fine-grained data.

The authors identify a clustering of the population from an administrative data that feeds into a multi-class $M/G/c$ model, whose parameters are recovered from the data via the Wasserstein distance. This model then informs an analysis of the underlying system and the needs of the population under study.

The analyses used herein consider, in effect, all types of patient arrivals and how they 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 their arrival.

KEYWORDS
OR in health services; machine learning; queuing

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 an abundance of data. However, many such methods rely 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 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. This work utilises a dataset of patients presenting

CONTACT Henry Wilde. Email: henrydavidwilde@gmail.com

chronic obstructive pulmonary disease (COPD), and demonstrates how insights can be identified by extracting information from administrative data. COPD is a condition of particular interest to population health research 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.

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 introduces the paper and provides a literature review, followed by an overview of the case study dataset and its clustering;
- Section 2 describes the queuing model and the estimation of its parameters;
- Section 3 presents several what-if scenarios with insight provided by the model parameterisation and the clustering;
- Section 4 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 in text or infographics; for instance, in: (Vuik, Mayer, & Darzi, 2016a, 2016b; Yan et al., 2019; Yoon, Goh, Kwan, Thumboo, & Low, 2020).

The review identified three 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, but groups their movements through the 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, & Low, 2018).

LCA is a statistical, model-based method used to identify groups (latent classes) in data by relating observations to some unobserved, categorical attribute. The discovered relations allow the observations to be separated into classes according to their maximum likelihood class membership (Hagenaars, 2002; Lazarsfeld & Henry, 1968).

This method has proved useful in the study of comorbidity patterns, as in Kuwornu, Lix, and Shooshtari (2014) or 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. Clustering methods are varied, but the common theme is to maximise homogeneity within, and heterogeneity between, each cluster (Everitt, Landau, Leese, & Stahl, 2011). Of these methods, the k-means paradigm is the most popular form in healthcare modelling literature. Some recent examples include: Elbattah and Molloy (2017); Haraty, Dimishkieh, and Masud (2015); Ogbuabor and Ugwoke (2018); Santhi, Bhaskaran, et al. (2010); Silitonga (2018); Vuik et al. (2016b). The k-means method iteratively partitions numerical data into $k \in \mathbb{N}$ distinct parts where k is fixed a priori, according to a heterogeneity function. This method's popularity is likely due to its simplicity, scalability, and that its implementations are concise (Olafsson, Li, & Wu, 2008; Wu & Kumar, 2009).

In addition to k-means, hierarchical clustering methods have proven useful in healthcare applications. In Vuik et al. (2016b), hierarchical clustering is used to identify a suitable number of patient clusters. Likewise, hierarchical clustering has been used to profile broader healthcare metrics such as patient utilisation patterns (Zayas et al., 2016) or mapping out effective leadership models Hargett et al. (2017). Also, supervised hierarchical segmentation methods such as classification and regression trees have been used where an existing, well-defined, label is of particular significance Harper and Winslett (2006); Kumar and Anjomshoa (2019).

1.1.2. Queuing models

Since the seminal works of Erlang (1917, 1920) established the core concepts of queuing theory, its application to real services has become abundant, including the healthcare service. Queuing theory is a mature discipline with many facets that extend beyond the needs of this manuscript. Comprehensive and informative introductions to queues and their simulation may be found in (Bhat, 2015; Shortle, Thompson, Gross, & Harris, 2018; Stewart, 2009).

Applying these models to healthcare settings can reveal many aspects of the underlying system. 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. More modern works Bittencourt, Verter, and Yalovsky (2018); Palvannan and Teow (2012); Pinto, de Campos, Perpétuo, and Ribeiro (2014) consider more detailed datasets to build their models. Moreover, model outputs are catered towards being actionable—as is the prerogative of operational research. For instance, Pinto et al. (2014) devises categorisations for hospital beds and arrivals. 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 theoretic models, queuing research has expanded to include computer simulation models. Simulating queues (or networks thereof) captures the stochastic nuances of hospital systems better than their theoretic counterparts. Example areas include the construction and simulation of Markov processes via process mining (Arnolds & Gartner, 2018; Rebuge & Ferreira, 2012), multi-class queuing networks (Cochran & Roche, 2009), and patient flow (Bhattacharjee & Ray, 2014).

There are numerous tools available for simulating queues, but few address core issues

like reproducibility. Dagkakis and Heavey (2016) provides a review on this subject. 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 the necessary technical expertise. However, they can foster poor simulation practices Bell and O'Keefe (1987). 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.

Reproducibility 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 these issues as how they are used involves less uncertainty and requires more rigour than 'drag-and-drop' software.

The simulation framework of choice for this manuscript is the discrete event simulation library, Ciw (Palmer, Knight, Harper, & Hawa, 2019). Ciw is written in Python, and is a well-developed piece of open-source software, adhering to best practices in research software development such as those set out in Benureau and Rougier (2018) and Jiménez et al. (2017). In Palmer et al. (2019), the authors stress how ensuring sustainable and reproducible simulation work are at the core of their development process.

1.1.3. Handling incomplete queue data

As discussed throughout this paper, the data used in this work is less detailed than in 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.

It is often the case that, 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. 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, the authors identify two articles as being applied to healthcare: Mohammadi and Salehi-Rad (2012); Yom-Tov and Mandelbaum (2014). Both works consider customers who can re-enter services during their time in the 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 filtering out those customers seeking to be serviced again. 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 modern queuing research. However, the scope is primarily focused on theoretic approximations rather than by simulation. Djabali, Rabta, and Aissani (2018) and Goldenshluger (2016) are two such works that attempt 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 summary of 5,231 patients presenting COPD from February 2011 through March 2019, covering 10,861 hospital spells. A patient 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. An episode is defined as any continuous period of care provided by the same consultant. Figure 1 contains an example of the relationship between episodes and spells.

[Figure 1 about here.]

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 conditions (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 another administrative dataset;
- 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.

The attributes included in the clustering encompass 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, and the engineered attributes (not including age or costs due to a lack of coverage).

With these attributes selected, a clustering algorithm must be chosen. Two crit-

ical specifications are that it must handle mixed-type data and be interpretable by stakeholders. The authors of Jahangirian, Borsci, Shah, and Taylor (2015) present an analysis of the factors resulting in a low level of engagement from stakeholders with healthcare simulation work. The key findings indicate that complexity and communication are the limiting factors for stakeholders, so the onus rests with researchers to make their models informative, effective and transferable.

Given these constraints, the k-prototypes algorithm is a strong candidate. Presented in Huang (1997) alongside the k-modes algorithm for categorical clustering, the k-prototypes algorithm is a mixed-type extension to the k-modes and k-means algorithms. In effect, the k-prototypes algorithm separates the dataset it acts on into its numeric and categorical attributes before applying k-means and k-modes on the respective parts. The cost functions for each of these parts are then combined to give a single cost function according to a weight, $\gamma \in \mathbb{R}$. This weight is also used to define the dissimilarity between two points, X and Y, in a dataset:

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

The choice of γ is of particular importance as it balances the contribution of each data type to the objective function. The seminal work Huang (1997) investigated the effect of various γ values, and determined that a sensible, robust value for γ is the average of the standard deviations for the numeric attributes. The analysis for 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 Satopaa, Albrecht, Irwin, and Raghavan (2011) was used with a range of potential values for k from two to ten. This range was based on what may be considered feasibly informative to stakeholders. This process revealed an optimal value for k of four, but both three and five clusters were considered; both were eliminated due to a lack of separation between the cluster characteristics.

[Table 1 about here.]

The dataset studied here is confidential, but a synthetic analogue illustrating the clustering has been archived under `DOI:10.5281/zenodo.3908167`. Table 1 provides a summary of the dataset and its clustering. Note that a negative length of stay indicates that the patient had passed away prior to arriving at the hospital; 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, insights can be gained about the segments identified by the clustering. For instance, the needs of each cluster can be summarised succinctly:

- Cluster 0 represents those spells with *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 costs.

- Cluster 2 represents the majority of spells and those where *resource requirements and clinical complexities are minimal*; these spells are the shortest, and the patients present with fewer diagnoses and a lower median CCI than any other cluster. Also, 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 but perhaps most critical section of the population: spells with *high complexity and high resource needs*. The patients within Cluster 3 are the oldest 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 concurrent diagnoses.

Figures 2 through 6 show the distributions for some of the 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, it has been included to highlight the value added by segmenting the population without such a prescriptive framework.

[Figure 2 about here.]

[Figure 3 about here.]

Figure 2 shows the length of stay distributions as histograms. Figure 2a demonstrates the different bed resource requirements well for each cluster in that their differences are not only a matter of varying means and ranges, but entirely different shapes to their distributions. All the distributions are 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 similar. The exception is the spells with no COPD intervention, where binning could not improve the visualisation due to the spread in their lengths of stay.

The same conclusions can be drawn from Figure 3 about costs; there are distinct patterns between the clusters, 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 immediate difference in the distribution of costs when splitting by intervention.

[Figure 4 about here.]

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.

[Figure 5 about here.]

[Figure 6 about here.]

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, the clinical complexity of each cluster can be captured better than with Table 1 alone. In Figure 5a, 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 to the population but is strongly biased towards patients presenting only COPD; Cluster 3 has the closest-to-uniform 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 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 align with the high rates of LTCs. Secondly, the vast majority of spells in Cluster 3 present with at least nine ICDs suggesting a wide range of conditions and comorbidities beyond the LTCs used to calculate CCI.

Conversely, little can be drawn from the same figures for the interventions (Figures 5b and 6b). One thing of note is that patients receiving both interventions (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.

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 identified in this section. However, to glean further insights, some parameters of the queuing model must be recovered from the data. The following section describes how these parameters are derived using the dataset at hand.

2. Constructing the queuing model

Following on from recent literature (Steins & Walther, 2013; Williams et al., 2015), this work employs a single node queue to model a hypothetical ward of patients presenting COPD. Additionally, the segmentation found in Section 1.2 provides a set of classes in the queue.

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 in Figure 2a, the lengths of stay require shifting prior to fitting an exponential distribution. As such, this work employs a $M/G/c$ queue with shifted exponential service time distributions for each cluster. Figure 7 provides a diagrammatic depiction of the process described in the remainder of this section to recover and simulate this $M/G/c$ queue.

[Figure 7 about here.]

2.1. Deriving the model parameters

Under the proposed $M/G/c$ model, the following assumptions are made:

(1) Inter-arrival and service times of patients are each exponentially distributed with some mean and a 'shift' defined in (2). 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. A queue capacity of zero is used in Williams et al. (2015) under the assumption that any surplus arrivals would be sent to another 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 rather than some priority framework.

Each group of patients has its arrival distribution, the parameter of which is the reciprocal of the mean inter-arrival time for that group. This parameter is denoted λ_l for each cluster l.

Like arrivals, each group of patients has its service time distribution. As noted earlier in this section, the lengths of stay must be shifted before they can be used to derive the service time 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:

$$\widehat{T}_l := \{t - m_l : t \in T_l\} \tag{2}$$

An exponential distribution may be fitted to these shifted system times by using their mean, denoted by $\frac{1}{\phi_l}$. 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:

$$\mu_l = p_l \phi_l \tag{3}$$

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

With these definitions, the service times for cluster l, denoted S_l, are 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:

$$f(s) = \begin{cases} \mu_l e^{-\mu_l(s-m_l)} & \text{if } s \geq m_l \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Therefore, the mean service time for spells in cluster l is given by:

$$\mathbb{E}(S_l) = \int_{m_l}^{\infty} \mu_l s e^{-\mu_l(s-m_l)} \mathrm{d}s = m_l + \frac{1}{\mu_l} \tag{5}$$

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 the proposed model allows for multiple classes and the shift terms are not the same for each cluster, the model does not have Markovian service times and is described as a $M/G/c$ model.

2.2. 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, 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 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' distance (Vaserstein, 1969). The Wasserstein metric satisfies the definition of a mathematical metric and takes 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. Formally, the Wasserstein distance between two probability distributions U and V is defined as:

$$W(U,V) = \int_0^1 \left|F^{-1}(t) - G^{-1}(t)\right| \mathrm{d}t \tag{6}$$

Here, F and G are the cumulative density functions of U and V, respectively. A proof of (6) 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. Specifically, 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:

$$(c^*, p^*) = \arg\min_{c,p} \left\{\max\left\{W\left(T_{c,p}, T\right)\right\}\right\} \tag{7}$$

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 in 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 perform 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 one year each, leaving two years of simulated data from each repetition.

[Figure 8 about here.]

The results of this parameter sweep are summarised in Figure 8. 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 as the differences in the quality of the fits are stark. In the best case the fit is uncanny, whereas the median case shows a distribution that inflates short-stay patients despite an otherwise good fit. Meanwhile, Figure 8c 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 from a disproportionate amount of long-stay patients. Table 2 reinforces these results numerically, showing a precise fit everywhere by the best parameter set.

[Table 2 about here.]

In this section, the identified clustering enriched the overall queuing model and was used to recover the parameters for several classes within that. 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 parameterisation of the queue introduced in Section 2. 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. 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 model parameterisation, and its assumptions, the effects on the chosen metrics in each scenario are in relative terms with respect to the base case—i.e., the results from the best parameter set recorded in Table 2. In particular, the data in each scenario is scaled by the corresponding median base case value, meaning that a metric having a value of 1 is 'normal'.

As mentioned in Section 1, the source code used throughout this manuscript has been archived online under `DOI:10.5281/zenodo.4457902`. Also, the datasets generated from the simulations in this section, and the parameter sweep from Section 2, have been archived online under `DOI:10.5281/zenodo.4457808`.

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. 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 cluster l, denoted $\hat{\lambda}_l$, is given by:

$$\hat{\lambda}_l = \sigma \lambda_l \tag{8}$$

[Figure 9 about here.]

Figure 9 shows the effect of changing patient arrivals on the relative system metrics for values of σ from 0.5 to 2.0 at a precision of 1.0×10^{-2}. Each plot in the figure (and the subsequent figures in this section) shows the median and interquartile range (IQR) of each relative metric. 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).

The indications of these plots align with what 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 9a, 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 reaches orders of magnitude above the norm.

However, Figure 9b shows that the system resources reach their worst case near to the start of that spike in relative system times (at $\sigma \approx 1.3$). That is, the median server utilisation hits its maximum at this point, and the variation in server utilisation disappears entirely. In this scenario, the servers are constantly active. 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.

3.2. Changes to resource availability

As 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 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 10 about here.]

Figure 10 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 bolster the claim from Section 3.1 that there is some room to manoeuvre where the system runs as normal. However, pressing on

those boundaries results in massive changes to both resource requirements and system times.

In Figure 10a, 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 resources (by merely being in a hospital) regardless of their condition, putting added strains on the system. Figure 10b 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 10b 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 with resources. 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.

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 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:

$$\hat{\lambda}_l = (1 - \delta)\,\lambda_l \quad \text{and} \quad \hat{\lambda}_m = \delta\lambda_l + \lambda_m \tag{9}$$

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 11 and 12 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}.

[Figure 11 about here.]

[Figure 12 about here.]

Considering Figure 11, 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.

[Table 3 about here.]

[Table 4 about here.]

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 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.

4. 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. 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 Section 3—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.

Benureau, F. C. Y., & Rougier, N. P. (2018). Re-run, repeat, reproduce, reuse, replicate: Transforming code into scientific contributions. *Frontiers in Neuroinformatics*, *11*.

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.

Elbattah, M., & Molloy, O. (2017). Clustering-aided approach for predicting patient outcomes with application to elderly healthcare in Ireland. In *Aaai workshops.*

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.

Haraty, R. A., Dimishkieh, M., & Masud, M. (2015). An enhanced k-means clustering algorithm for pattern discovery in healthcare data. *International Journal of Distributed Sensor Networks*, *11*(6), 615740.

Hargett, C., Doty, J., Hauck, J., Webb, A., Cook, S., Tsipis, N., ... Taylor, D. (2017). Developing a model for effective leadership in healthcare: a concept mapping approach. *Journal of Healthcare Leadership*, *Volume 9*, 69–78.

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).

Jahangirian, M., Borsci, S., Shah, S. G. S., & Taylor, S. J. E. (2015). Causal factors of low stakeholder engagement: a survey of expert opinions in the context of healthcare simulation projects. *SIMULATION*, *91*(6), 511–526.

Jiménez, R. C., Kuzak, M., Alhamdoosh, M., Barker, M., Batut, B., Borg, M., ... Crouch, S. (2017). Four simple recommendations to encourage best practices in research software. *F1000Research*, *6*, ELIXIR-876.

Komashie, A., Mousavi, A., Clarkson, P. J., & Young, T. (2015). An integrated model of pa-

tient 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.

Kumar, A., & Anjomshoa, H. (2019). A two-stage model to predict surgical patients' lengths of stay from an electronic patient database. *IEEE Journal of Biomedical and Health Informatics*, *23*(2), 848-856.

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.

Ogbuabor, G., & Ugwoke, F. (2018). Clustering algorithm for a healthcare dataset using silhouette score value. *International Journal of Computer Science & Information Technology*, *10*(2), 27–37.

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.

Santhi, P., Bhaskaran, V. M., et al. (2010). Performance of clustering algorithms in healthcare database. *International Journal for Advances in Computer Science*, *2*(1), 26–31.

Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a 'kneedle' in a haystack: Detecting knee points in system behavior. In *Proceedings of the 2011 31st international conference on distributed computing systems workshops* (pp. 166–171).

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.

Silitonga, P. (2018). Clustering of patient disease data by using k-means clustering. *International Journal of Computer Science and Information Security*, *15*(7), 219–221.

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 Peredačhi 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.

Zayas, C. E., He, Z., Yuan, J., Maldonado-Molina, M., Hogan, W., Modave, F., ... Bian, J. (2016). Examining healthcare utilization patterns of elderly middle-aged adults in the United States. In *Proceedings of the... international florida ai research society conference. florida ai research symposium* (Vol. 2016, p. 361).

		Cluster 0	1	2	3	Population
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

		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

Origin	δ Destination	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
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

Origin	δ Destination	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
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

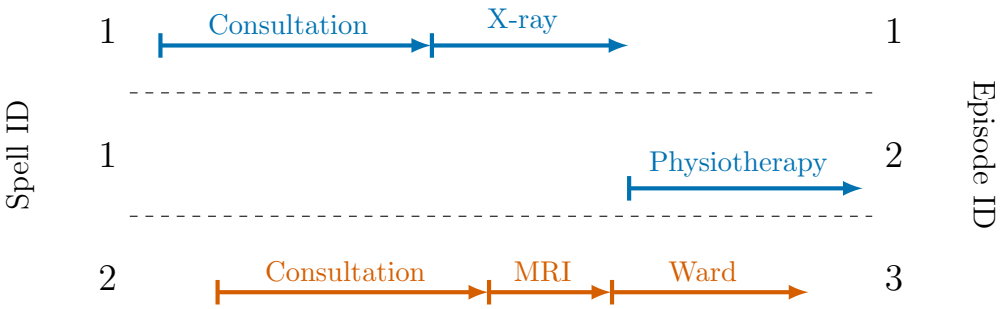
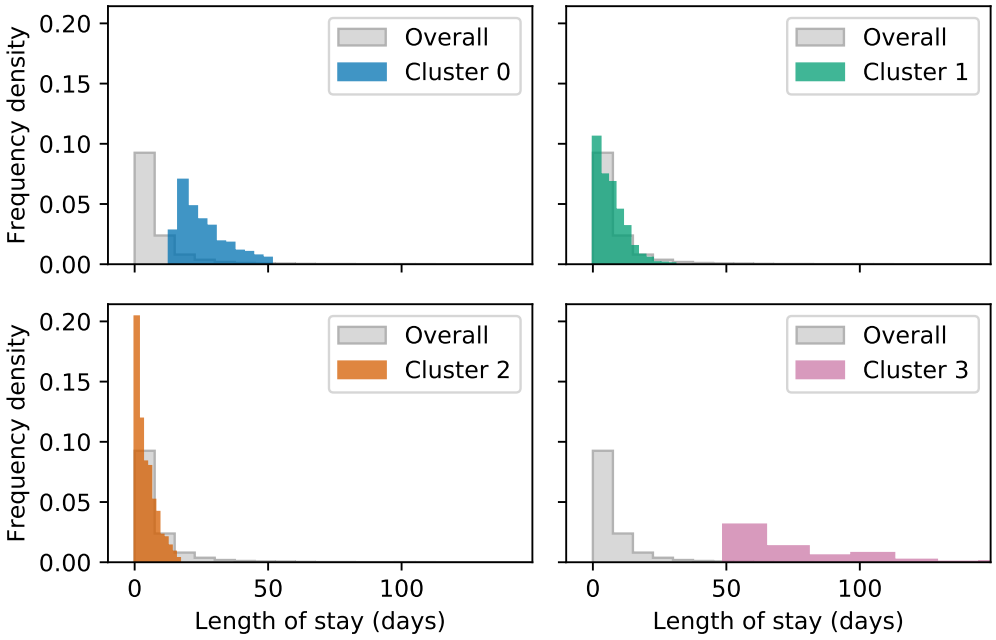
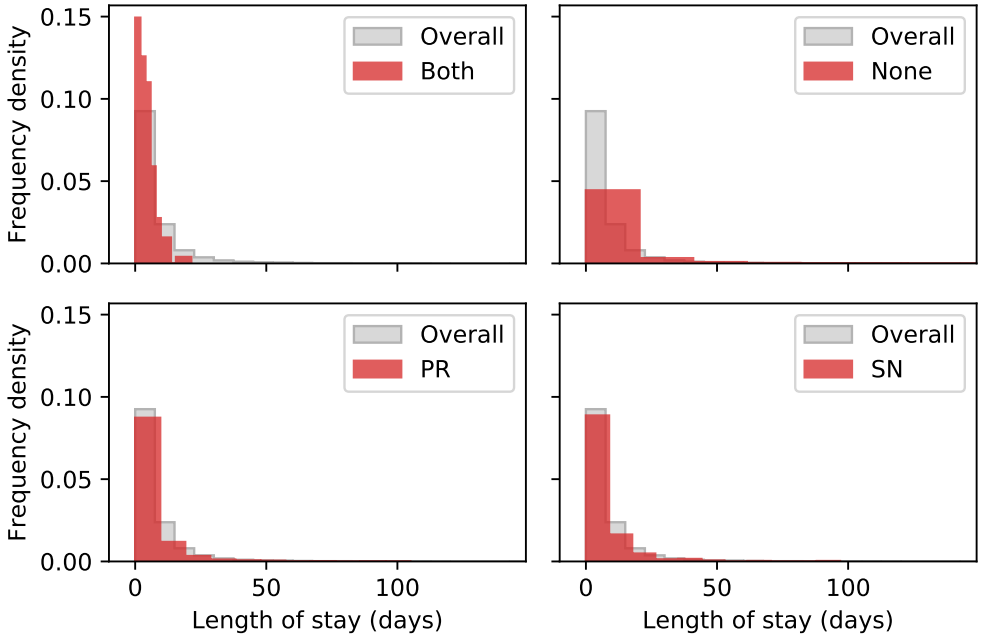


Figure 1. A Gantt chart of two patient spells across three episodes

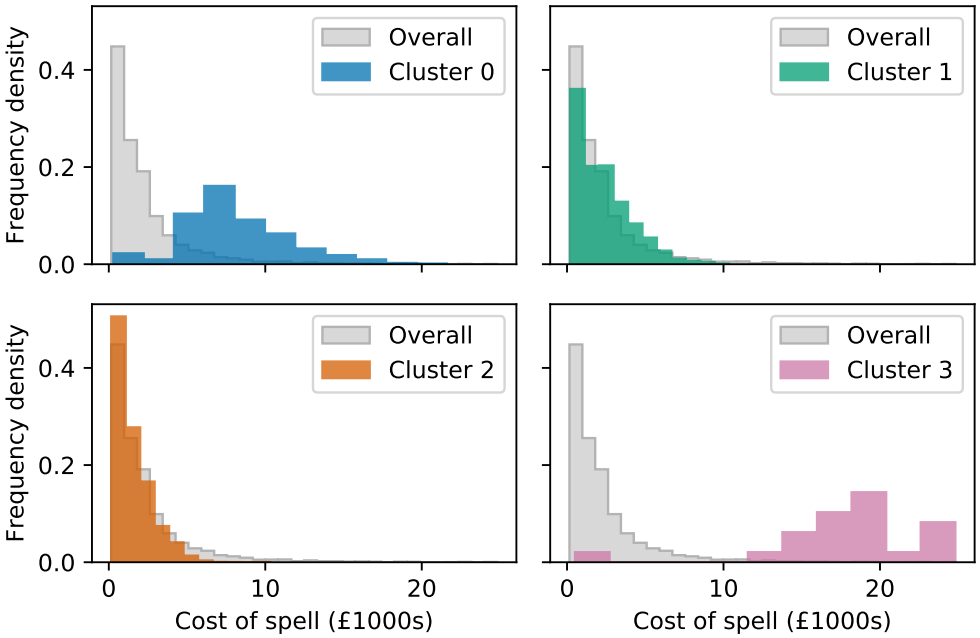


(a)

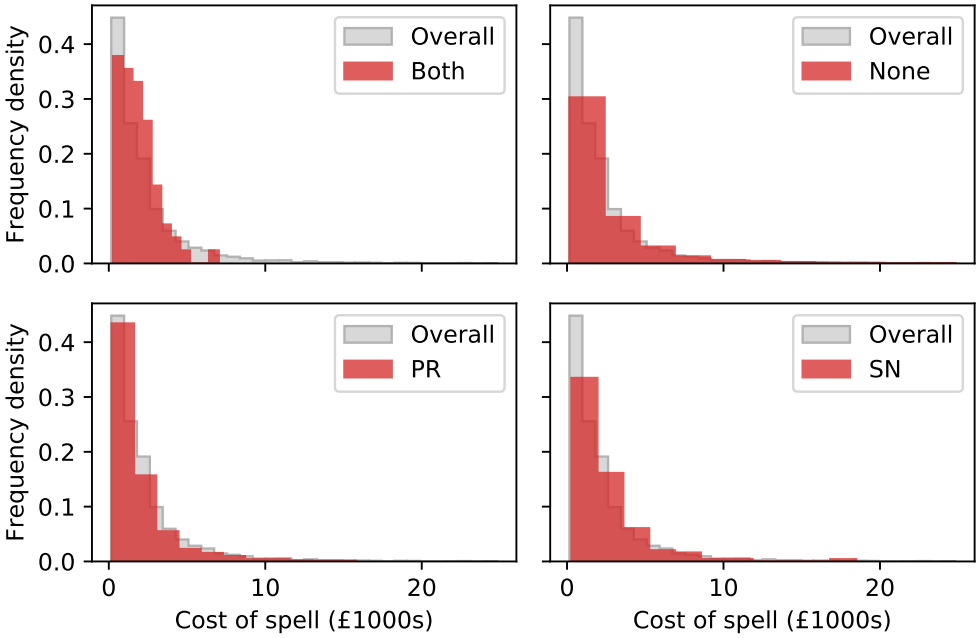


(b)

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

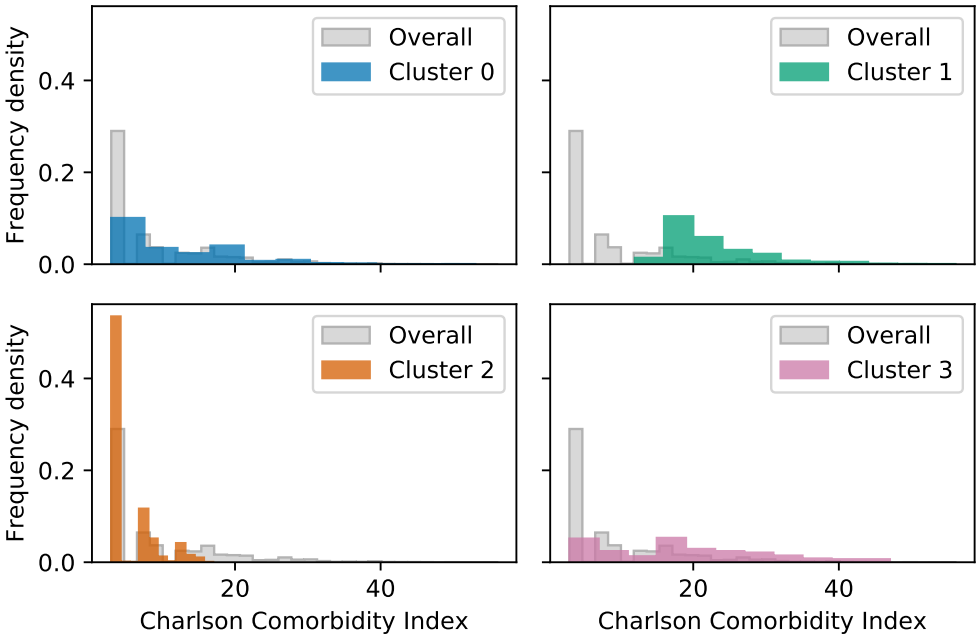


(a)

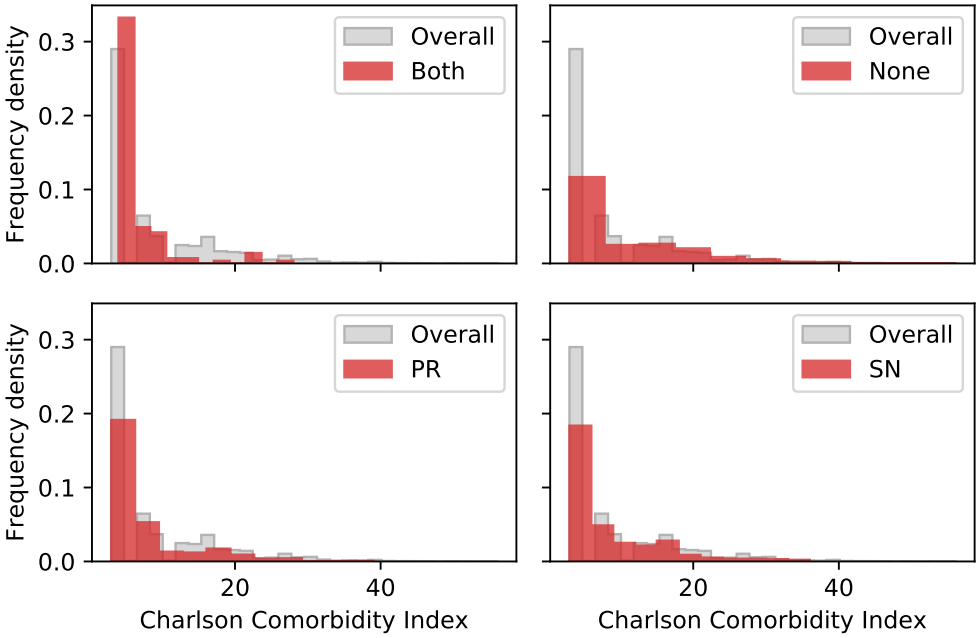


(b)

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

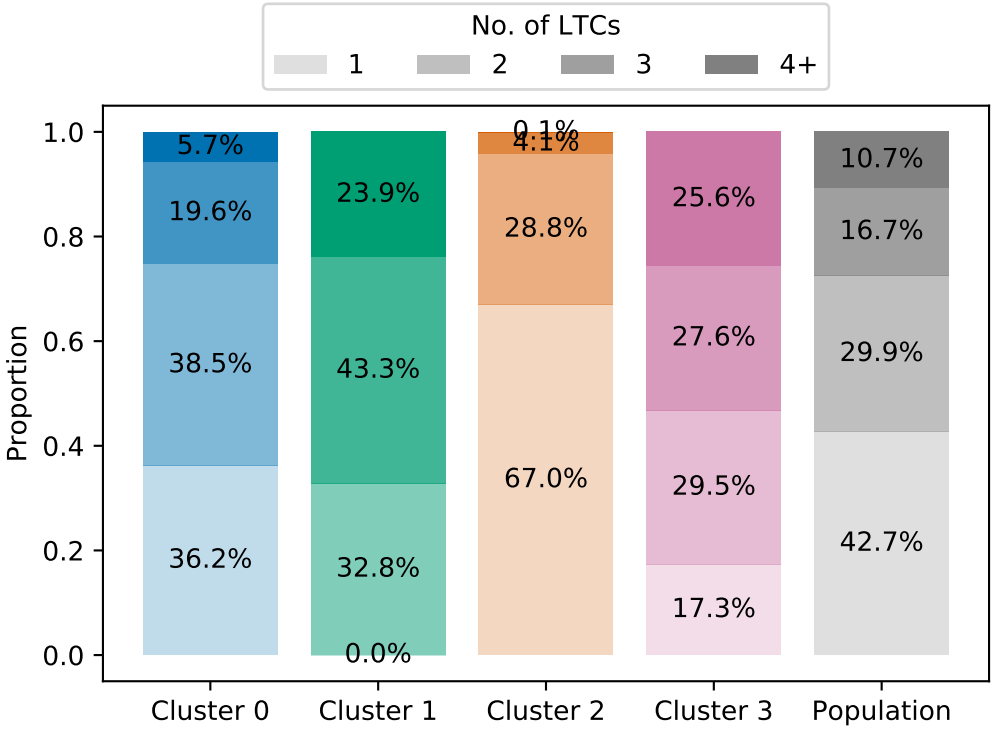


(a)

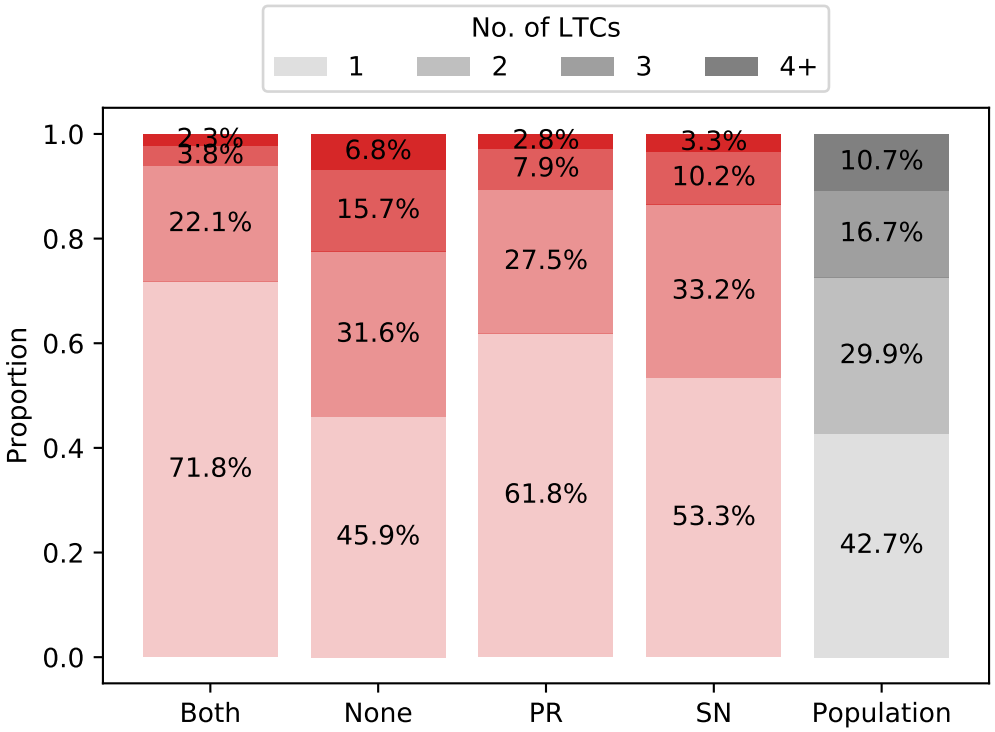


(b)

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

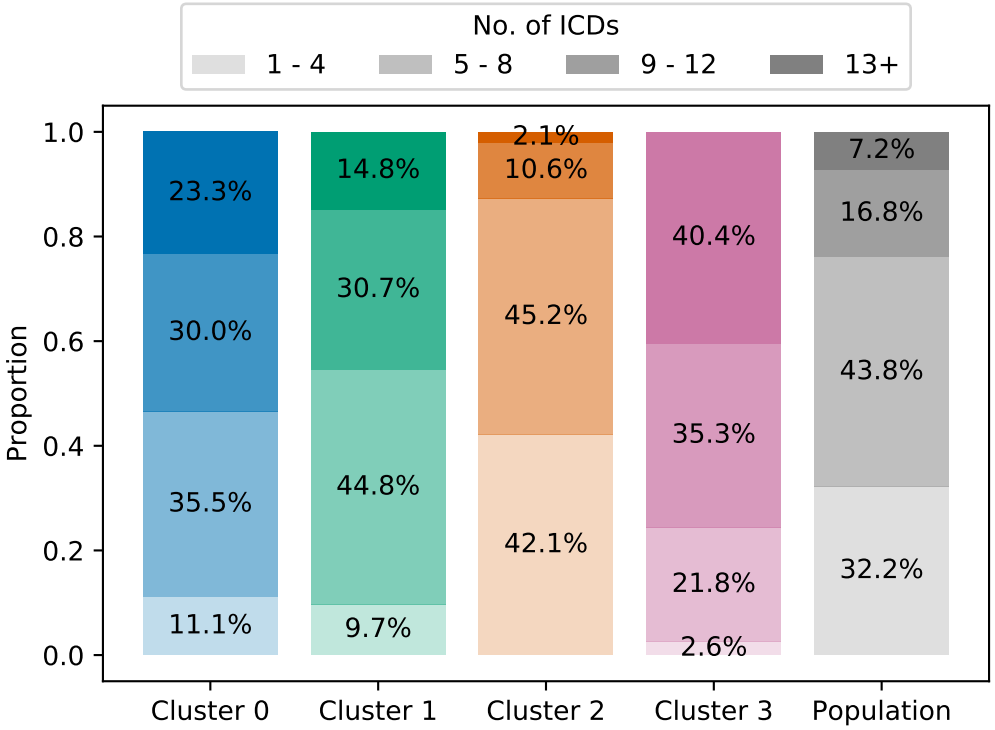


(a)

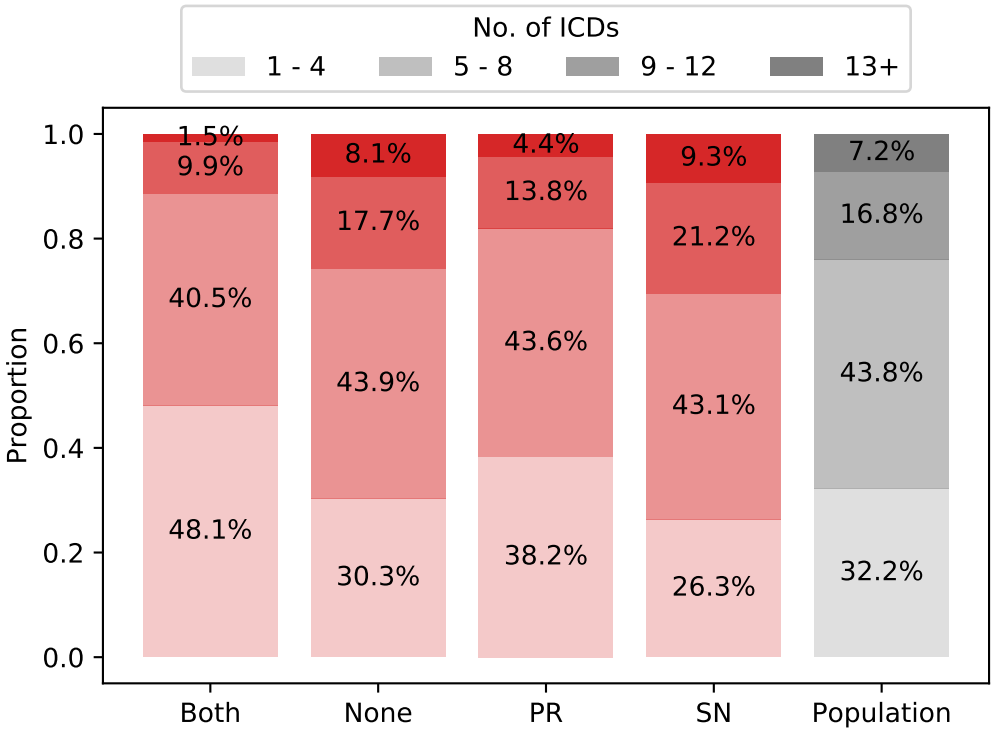


(b)

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

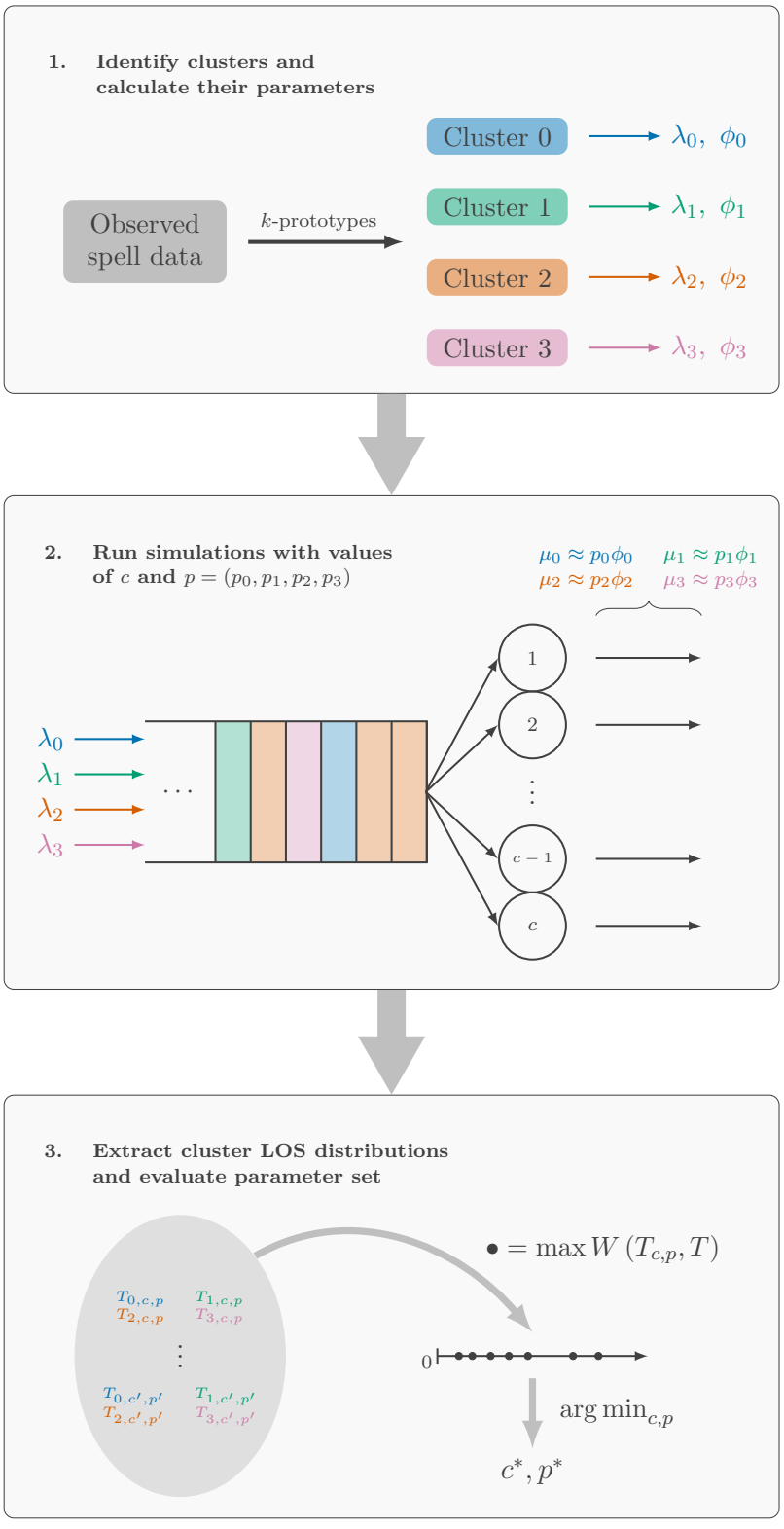
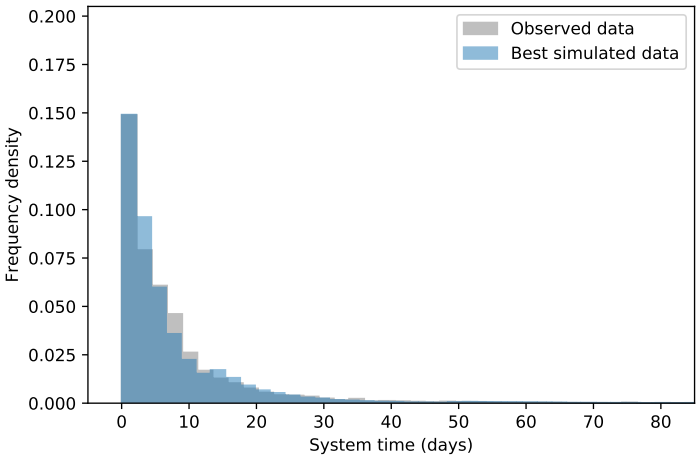
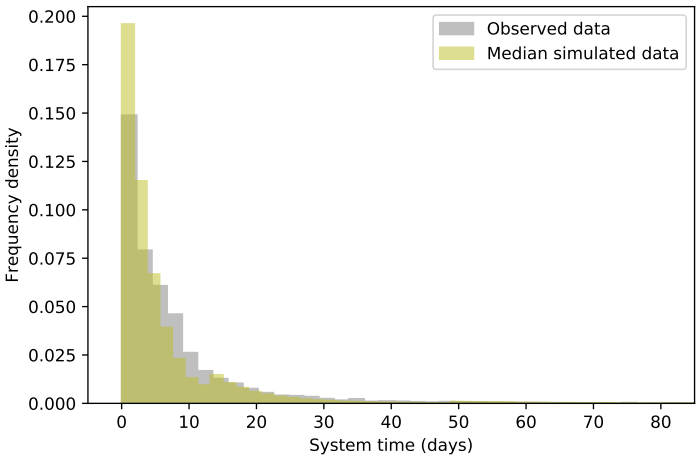


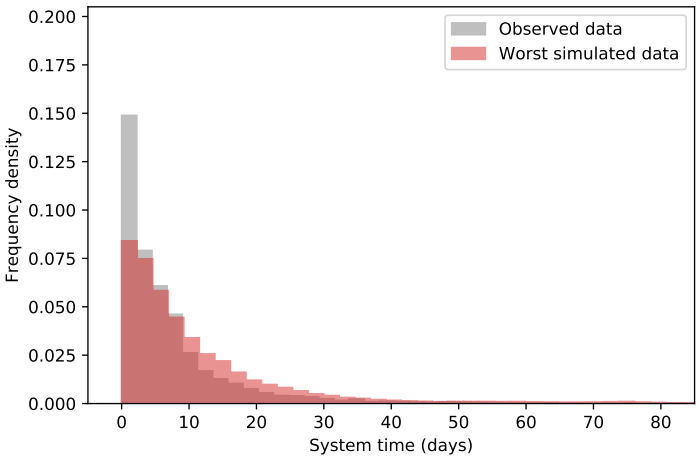
Figure 7. A diagrammatic depiction of the queuing parameter recovery process



(a)

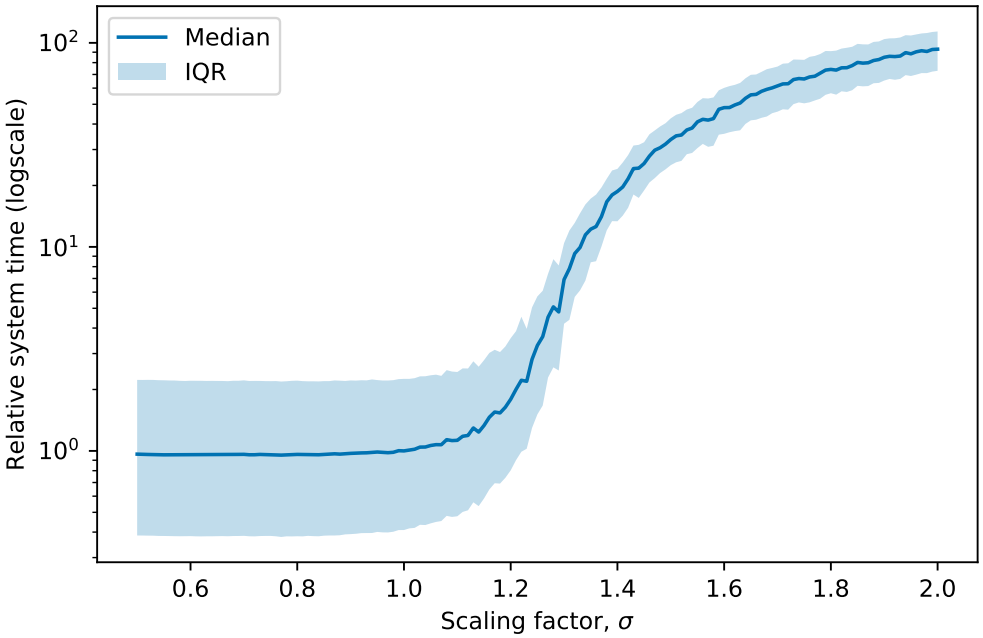


(b)

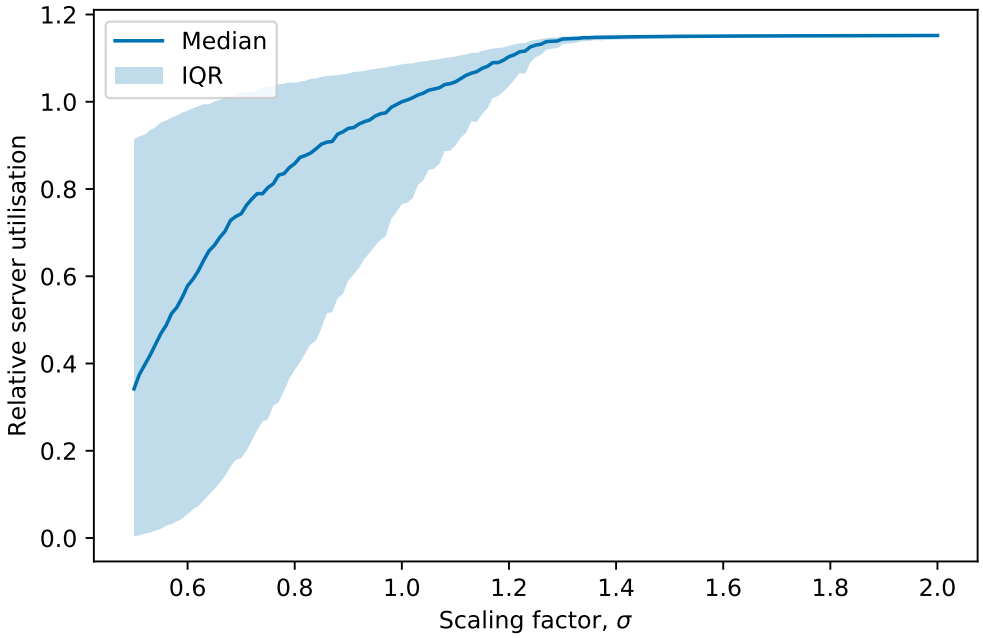


(c)

Figure 8. Histograms of the observed LOS data and the (a) best-simulated, (b) median-simulated, and (c) worst-simulated LOS data.

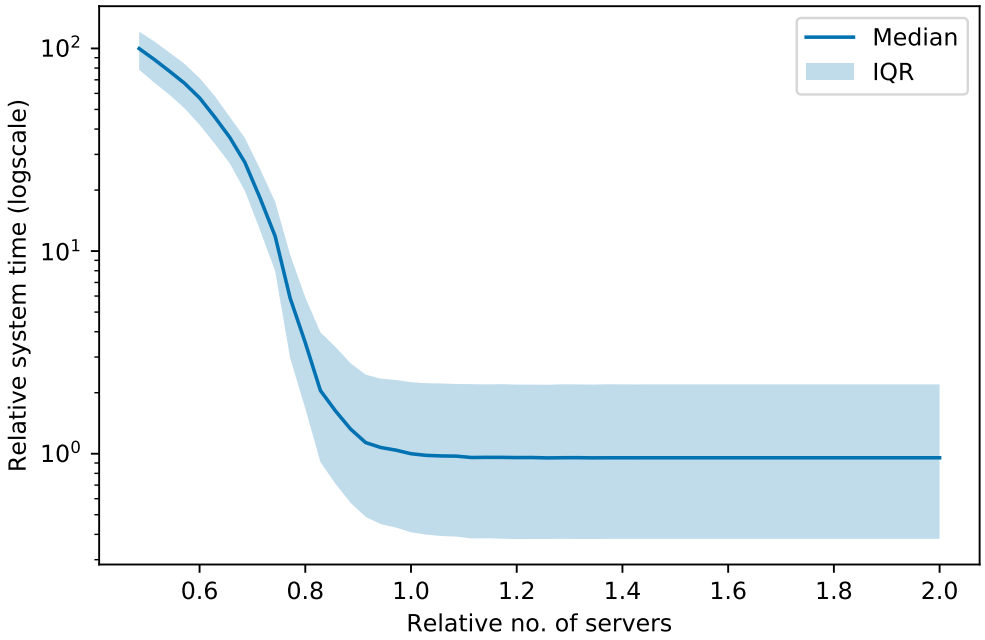


(a)

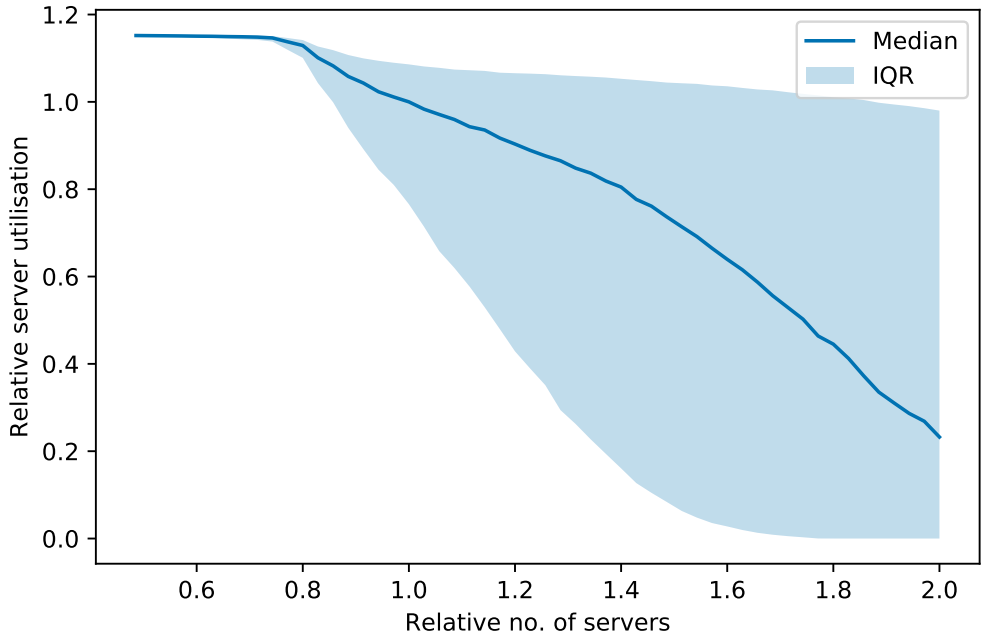


(b)

Figure 9. Plots of σ against relative (a) system time and (b) server utilisation



(a)



(b)

Figure 10. Plots of the relative number of servers against relative (a) system time and (b) server utilisation

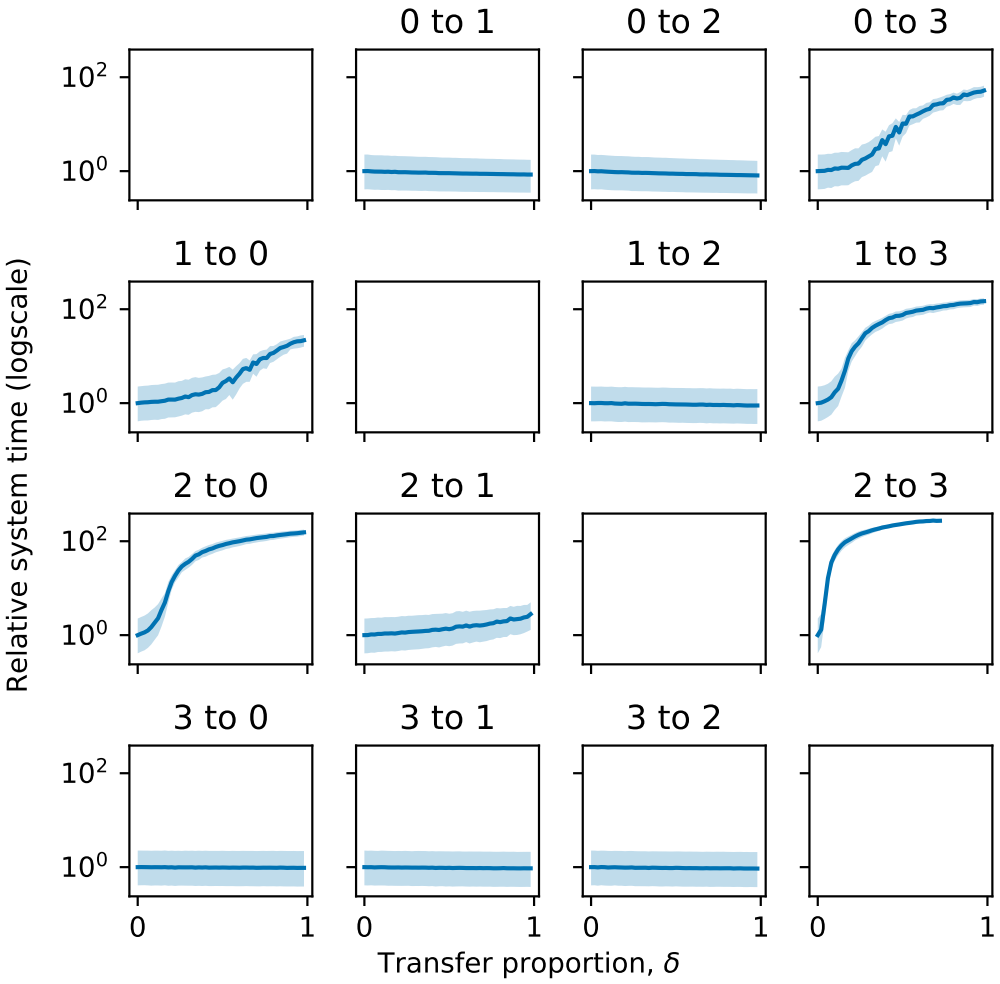


Figure 11. Plots of proportions of each cluster moving to another against relative system time

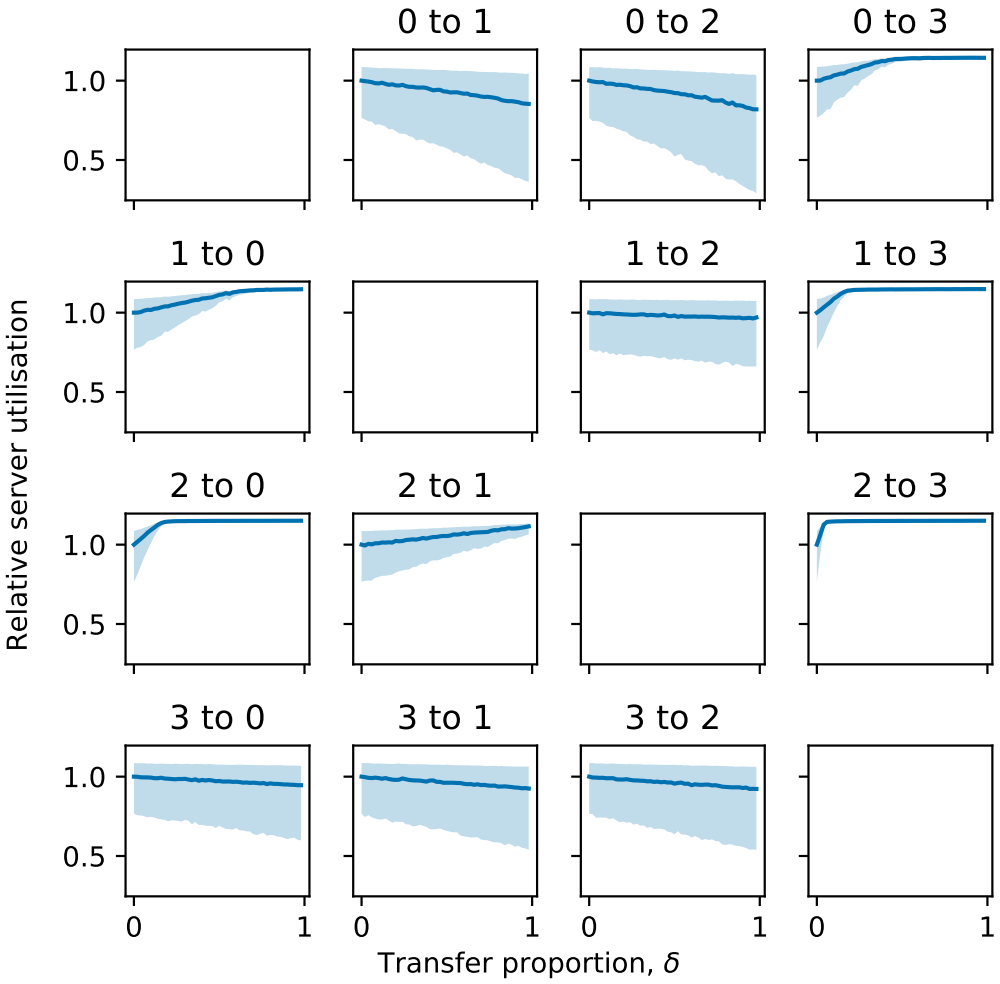


Figure 12. Plots of proportions of each cluster moving to another on relative server utilisation