# Understanding COPD patients in the hospital system via administrative data

Henry Wilde, Vincent Knight, Jonathan Gillard

**Abstract**

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

## 1 Introduction

Population health research is becoming 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. A vital part of such research is to better understand the healthcare needs and behaviours of a population, and it can be beneficial to find an appropriate segmentation of that population; such a segmentation allows for finer-grained analysis of groups in the population that share some form of homogeneity. One commonly used method for such patient-centred analysis is that of patient flow and their interaction with the healthcare system.

However, this process relies heavily on detailed data — about both the system and the population within that system — which may limit research where sophisticated data pipelines are not yet in place. This work demonstrates how this issue may be overcome using administrative, spell-level hospital data to build a patient clustering that feeds into a multi-class queuing model. Specifically, this work examines patient records from the National Health Service (NHS) Wales Cwm Taf Morgannwg University Health Board (UHB) that present chronic obstructive pulmonary disease (COPD). COPD is of particular interest to Cwm Taf Morgannwg UHB as the condition is known to often present as a comorbidity in patients [12] and it was found that they had the highest prevalence of the condition across all the Welsh health boards in an internal report by NHS Wales.

The remainder of the paper is structured as follows: Section 1 provides a literature review, and an overview of the data and its clustering; Section **??** describes the queuing model used and the estimation of its parameters; Section 3 presents a number of what-if scenarios with insight provided by the model parameterisation and the clustering; Section 4 concludes the paper. Although the data is confidential and may not be published, the source code used in this paper is available online at `https://github.com/daffidwilde/copd-paper`.

## 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 principal topics of segmentation analysis, queuing models applied to hospital systems, and the handling of missing or incomplete data for such queues.

### 1.1.1 Segmentation analysis

Segmentation analysis allows for the targeted analysis of otherwise heterogeneous datasets and encompasses several techniques from operational research, statistics and machine learning. One of the most desirable qualities of this kind of analysis is the ability to glean and communicate simplified summaries of patient needs to stakeholders within a healthcare system [22, 28]. For instance, clinical profiling often forms part of the wider analysis where each segment can be summarised in a phrase or infographic [21, 26].

The review for this work identified three commonplace groups of patient characteristics used to segment a patient population: their system utilisation metrics, their clinical attributes and their pathway. The latter is not used to segment the patients directly but rather groups their movements through a healthcare system. This is typically done via process mining. [2] and [4] 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 with a number of techniques but recent works tend to use unsupervised methods, typically latent class analysis (LCA) or clustering [25].

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 categories, each corresponding to a latent class. The discovered relations are then used to separate the observations into latent classes according to their maximum

likelihood class membership [10, 15]. This method has proved useful in the study of comorbidity patterns as in [1, 14] 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 a labelling of its instances. However, clustering includes a wide variety of methods where the common theme is to maximise homogeneity within, and heterogeneity between, each cluster [8]. 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 [17, 24]. In addition to $k$-means, hierarchical clustering methods can be effective if a suitable number of parts cannot be found initially [21]. Although, supervised hierarchical segmentation methods such as classification and regression trees (as in [11]) have been used where an existing, well-defined label is of particular significance.

### 1.1.2 Queuing models

These are the principal queuing theory works by Erlang [6, 7]. Deadlock is an aspect of applied queuing theory of interest in recent literature [18]. The software used is Ciw [19].

### 1.1.3 Handling incomplete queue data

It is often the case that in practical situations where suitable data is not (immediately) available, further inquiry will stop in that particular line of research. Queuing models in healthcare settings appear to be such a case. [3] 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: [16, 27]. Both works are concerned with customers that can re-enter services during their time in the queuing system. This is particularly of value when considering the effect of unpredictable behaviour in intensive care units, for instance. [16] seeks to approximate service and re-service densities through a Bayesian approach and by separating out those customers seeking to be serviced again. On the other hand, [27] considers an extension to the $M/M/c$ queue with direct re-entries. The devised model is then used to determine resource requirements in two healthcare settings.

Aside from healthcare-specific works, the approximation of queue parameters has formed a part of relevant modern queuing research. However, the scope is largely focused on theoretic approximations rather than by simulation. [5, 9] 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.

## 1.2 Overview of the dataset and its clustering

The dataset used in this work was provided by the Cwm Taf Morgannwg UHB as part of an ongoing research project with the authors. The dataset contains a spell-level summary of 5,231 patients presenting COPD from February 2011 through March 2019 totalling 10,861 spells. The dataset is made up of instances each describing a patient spell with the following attributes:

- Personal identifiers and information, i.e. patient and spell ID numbers, and 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 spell is one of no intervention, pulmonary rehabilitation, specialist nursing, 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.
- Rank under the 2019 Welsh Index of Multiple Deprivation (WIMD) indicating relative deprivation of the postcode area the patient lives in.

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 generalised to their categories according to NHS documentation and counts for each category were attached.
- The number of COPD-related admissions in the last twelve months based on the associated patient ID number.

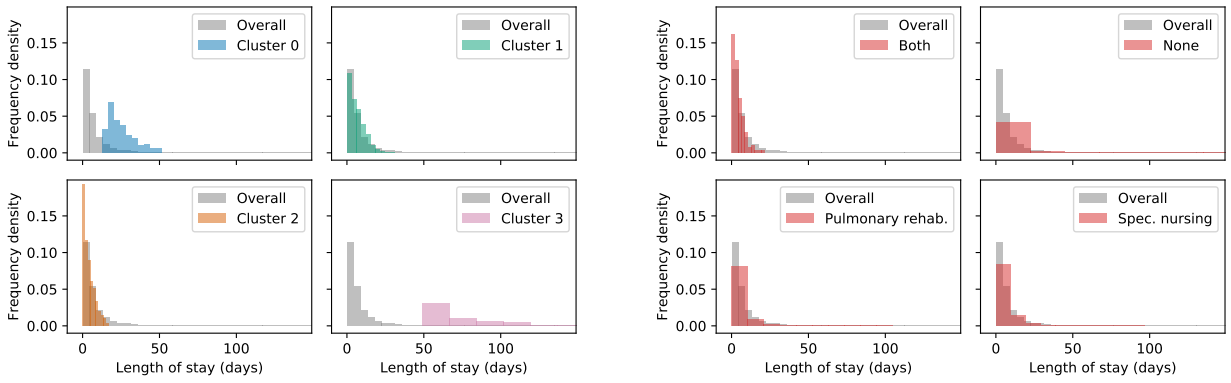



Figure 1: Historgrams for length of stay by (a) cluster and (b) intervention.

Due to a lack of information about the patients beyond their COPD-related admissions, the spells of the dataset were segmented. A variant of the $k$-means algorithm was used. This variant, 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. Both $k$-prototypes and $k$-modes were presented in [13].

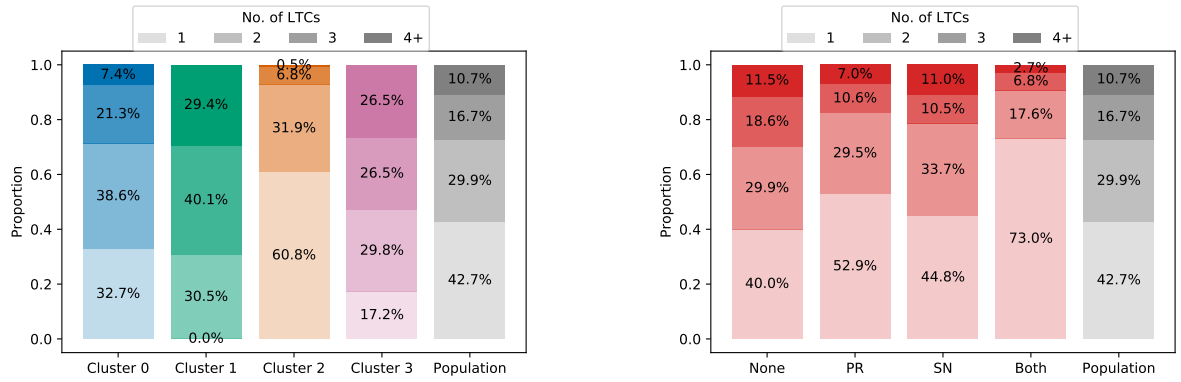
The attributes included in the clustering encompass both utilisation metrics and clinical attributes relating to the spell. They were as follows: the summative clinical path attributes, the CCI contributions and condition indicators, the WIMD rank, length of stay (LOS), COPD intervention status, and 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 [20] was used with a range of potential values for $k$ from 2 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 popular 'elbow method' for determining a number of clusters. This revealed an optimal value for $k$ of 4 but both 3 and 5 clusters were considered. Each case was eliminated due to a lack of clear separation in the characteristics of the clusters. Additionally, the initialisation method used for $k$-prototypes was that presented in [23] as it was found to give an improvement in the clustering over other initialisation methods.

A summary of the spells in each cluster, and the overall dataset (referred to as the population), is provided in Table 1.

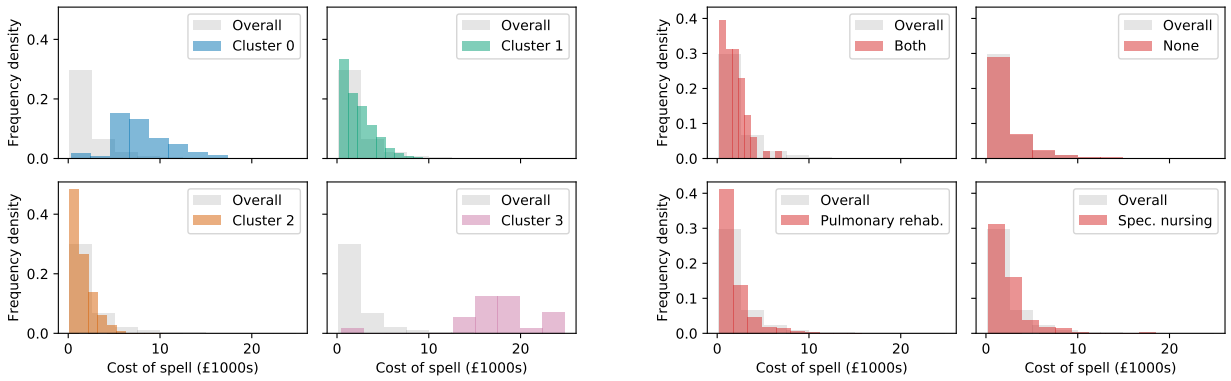
| | | Cluster | | | | Population |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | |
| **Characteristics** | **Mean spell cost, £** | 8083.69 | 2312.39 | 1509.22 | 17847.80 | 2280.54 |
| | **Mean age** | 76.09 | 76.66 | 70.55 | 81.70 | 72.22 |
| | **COPD adm. last year** | 2.20 | 1.97 | 1.88 | 2.08 | 1.93 |
| | **Minimum LOS** | 12.82 | 0.01 | 0.00 | 48.82 | 0.00 |
| | **Mean LOS** | 25.32 | 6.47 | 4.11 | 75.20 | 7.70 |
| | **Maximum LOS** | 51.36 | 30.86 | 16.94 | 224.93 | 224.93 |
| | **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, %** | 80.26 | 83.39 | 65.75 | 89.81 | 0.71 |
| | **Pulmonary rehab., %** | 15.77 | 13.41 | 27.96 | 8.92 | 0.24 |
| | **Spec. nursing, %** | 3.78 | 2.91 | 4.63 | 1.27 | 0.04 |
| | **Both, %** | 0.18 | 0.29 | 1.66 | 0.00 | 0.01 |
| **LTC prevalence** | **Pulmonary disease, %** | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| | **Diabetes, %** | 19.05 | 28.15 | 14.84 | 25.00 | 17.97 |
| | **AMI, %** | 13.85 | 22.94 | 8.76 | 16.03 | 12.10 |
| | **CHF, %** | 12.45 | 53.82 | 0.00 | 26.28 | 11.98 |
| | **Renal disease, %** | 7.53 | 19.55 | 1.92 | 17.95 | 6.11 |
| | **Cancer, %** | 7.62 | 12.24 | 2.93 | 10.90 | 5.30 |
| | **Dementia, %** | 6.88 | 21.27 | 0.00 | 26.92 | 5.17 |
| | **CVA, %** | 8.64 | 13.34 | 0.70 | 19.87 | 4.20 |
| | **PVD, %** | 4.37 | 7.70 | 2.27 | 5.77 | 3.57 |
| | **CTD, %** | 5.11 | 4.25 | 3.11 | 4.49 | 3.55 |
| | **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 |

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



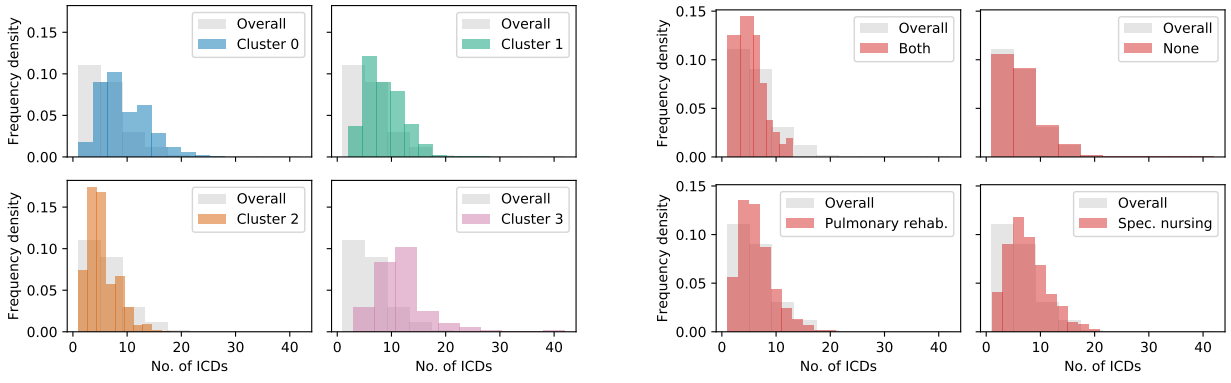

(a) (b)

Figure 2: Proportions of concurrent LTC counts presented by patients by (a) cluster and (b) intervention.



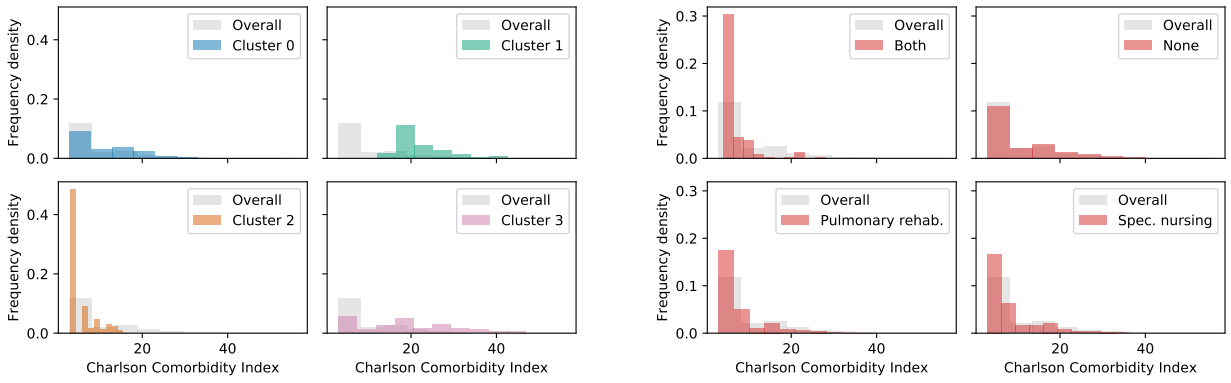

(a) (b)

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




(a) (b)

Figure 4: Histograms for number of ICDs by (a) cluster and (b) intervention.




(a) (b)

Figure 5: Histograms for CCI by (a) cluster and (b) intervention.

# 2 Estimating queuing parameters

Reiterate the objective of the paper — to model a COPD ward within a hospital — and draw attention to lack of fine-grain data. Lead into how this can be overcome with the Wasserstein distance (a lot of this has been written up in `nbs/wasserstein.ipynb`). A brief summary of how the parameter set is chosen and a nice image of the queue we are building. Close out the section with best and worst case parameter set plots.

# 3 Adjusting the queuing model

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

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

# 4 Conclusion

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

# References


[1] 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. 12, 2017.

[2] I. V. Arnolds and D. Gartner. Improving hospital layout planning through clinical pathway mining. *Annals of Operations Research*, 263:453 – 477, 2018. doi: 10.1007/s10479-017-2485-4.

[3] A. Asanjarani, Y. Nazarathy, and P. Pollett. Parameter and state estimation in queues and related stochastic models: A bibliography, 2017. URL `https://people.smp.uq.edu.au/PhilipPollett/papers/Qest/QEstAnnBib.pdf`.

[4] P. Delias, M. Doumpos, E. Grigoroudis, P. Manolitzas, and N. Matsatsinis. Supporting healthcare management decisions via robust clustering of event logs. *Knowledge-Based Systems*, 84:203 – 213, 2015. doi: 10.1016/j.knosys.2015.04.012.

[5] Y. Djabali, B. Rabta, and D. Aissani. 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, 06 2018. doi: 10.1504/IJMOR.2018.10005095.

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

[7] A. K. Erlang. Telephone waiting times. *Matematisk Tidsskrift, B*, 31:25, 1920.

[8] B. S. Everitt, S. Landau, M. Leese, and D. Stahl. *Cluster analysis*. John Wiley & Sons, 2011.

[9] A. Goldenshluger. Nonparametric estimation of the service time distribution in the M/G/$\infty$ queue. *Advances in Applied Probability*, 48(4):1117–1138, 2016. doi: 10.1017/apr.2016.67.

[10] J. A. Hagenaars. *Applied Latent Class Analysis*. Cambridge University Press, 2002. doi: 10.1017/CBO9780511499531.

[11] P. R. Harper and D. Winslett. Classification trees: A possible method for maternity risk grouping. *European Journal of Operational Research*, 169:146–156, 2006. doi: 10.1016/j.ejor.2004.05.014.

[12] S. Houben-Wilke, F. J. J. Triest, F. M. Franssen, D. J. Janssen, E. F. Wouters, and L. E. Vanfleteren. 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, 2019. doi: 10.15326/jcopdf.6.2.2018.0145.

[13] Z. Huang. Extensions to the $k$-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304, 1998. doi: 10.1023/A:1009769707641.

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

[15] P. F. Lazarsfeld and N. W. Henry. *Latent structure analysis*. Houghton Mifflin Co., 1968.

[16] A. Mohammadi and M. R. Salehi-Rad. Bayesian inference and prediction in an $M/G/1$ with optional second service. *Communications in Statistics - Simulation and Computation*, 41(3):419–435, 2012. doi: 10.1080/03610918.2011.588358.

[17] S. Olafsson, X. Li, and S. Wu. Operations research and data mining. *European Journal of Operational Research*, 187(3):1429 – 1448, 2008. doi: https://doi.org/10.1016/j.ejor.2006.09.023.

[18] G. I. Palmer, P. R. Harper, and V. A. Knight. Modelling deadlock in open restricted queueing networks. *European Journal of Operational Research*, 266(2):609 – 621, 2018. doi: 10.1016/j.ejor.2017.10.039.

[19] G. I. Palmer, V. A. Knight, P. R. Harper, and A. L. Hawa. Ciw: An open-source discrete event simulation library. *Journal of Simulation*, 13(1):68–82, 2019. doi: 10.1080/17477778.2018.1473909.

[20] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan. 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*, pages 166–171, 07 2011. doi: 10.1109/ICDCSW.2011.20.

[21] S. I. Vuik, E. K. Mayer, and A. Darzi. A quantitative evidence base for population health: Applying utilization-based cluster analysis to segment a patient population. *Population Health Metrics*, 14, 2016. doi: 10.1186/s12963-016-0115-z.

[22] S. I. Vuik, E. K. Mayer, and A. Darzi. Patient segmentation analysis offers significant benefits for integrated care and support. *Health Affairs*, 35(5):769–775, 2016. doi: 10.1377/hlthaff.2015.1311.

[23] H. Wilde, V. Knight, and J. Gillard. A novel initialisation based on hospital-resident assignment for the k-modes algorithm, 2020.

[24] X. Wu and V. Kumar. *The top ten algorithms in data mining*. CRC press, 2009.

[25] S. Yan, Y. H. Kwan, C. S. Tan, J. Thumboo, and L. L. Low. A systematic review of the clinical application of data-driven population segmentation analysis. *BMC Medical Research Methodology*, 18(121), 2018. doi: 10.1186/s12874-018-0584-9.

[26] S. Yan, B. J. J. Seng, Y. H. Kwan, C. S. Tan, J. H. M. Quah, J. Thumboo, and L. L. Low. Identifying heterogeneous health profiles of primary care utilizers and their differential healthcare utilization and mortality – a retrospective cohort study. *BMC Family Practice*, 20(54), 2019. doi: 10.1186/s12875-019-0939-2.

[27] G. B. Yom-Tov and A. Mandelbaum. Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2):283–299, 2014. doi: 10.1287/msom.2013.0474.

[28] S. Yoon, H. Goh, Y. H. Kwan, J. Thumboo, and L. L. Low. Identifying optimal indicators and purposes of population segmentation through engagement of key stakeholders: A qualitative study. *Health Res Policy Syst.*, 18(1):26, 2020. doi: 10.1186/s12961-019-0519-x.