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, i.e. a dataset that summarises patient spells. This data is then used to build a patient clustering that feeds into a multi-class queuing model. Specifically, this work examines records of patient spells 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 [15] 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 Although the data is confidential and may not be published, the source code used in this paper is available online at `https://github.com/daffidwilde/copd-paper`.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 [33, 39]. For instance, clinical profiling often forms part of the wider analysis where each segment can be summarised in a phrase or infographic [32, 37].

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. [1] and [6] 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 [36].

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 [13, 21]. This method has proved useful in the study of comorbidity patterns as in [19, 20] 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 [10]. 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 [25, 35]. In addition to k-means, hierarchical clustering methods can be effective if a suitable number of parts cannot be found initially [32]. Although, supervised hierarchical segmentation methods such as classification and regression trees (as in [14]) have been used where an existing, well-defined label is of particular significance.

1.1.2 Queuing models

Since the seminal works by Erlang [8, 9] established the core concepts of queuing theory, the application of queues and queuing networks to real services has become abundant including the healthcare service. By applying these models to healthcare settings, many aspects of the underlying system can be studied. A common area of study in healthcare settings is of service capacity. [22] is an early example of such work where acute bed capacity was determined using hospital occupancy data. Meanwhile, more modern works such as [27, 28] consider wider sources of data (where available) to build their queuing models. Moreover, the output of a model is catered more towards being actionable — as is the prerogative of operational research. For instance, [28] devises new categorisations for both hospital beds and arrivals that are informed by the queuing model. A further example is [18] where queuing models are used to measure and understand satisfaction amongst patients and staff.

In addition to these theoretic models, healthcare queuing research has expanded to include computer simulation models. The simulation of queues, or networks thereof, have the benefit of being able to easily capture the stochastic nuances of hospital systems over their theoretic counterparts. Example areas include the construction and simulation of Markov processes via process mining [1, 40], and patient flow [3]. Regardless of the advantages of simulation models, a prerequisite is reliable software with which to construct those simulations. A popular tool for building queues — both in industry and academia — is Simul8. This piece of software is based on processes and is highly visual which makes it attractive to organisations looking to implement queuing models without necessary technical expertise, including the NHS. [4] discusses the issues around operational research and simulation being taken up in the NHS despite the availability of intuitive software like Simul8. However, it does not address a core principle of good simulation work: reproducibility. The ability to reliably reproduce a set of results is a matter of great importance to scientific research but this remains an issue in simulation research generally [11]. When considering issues with reproducibility in scientific computing (simulation included), the buck often ends with the software used [17]. The use of well-developed, open source software can alleviate issues around reproducibility and reliability as the processes by which they are used involve less uncertainty and require more rigour than 'drag-and-drop' software. One example of such a piece of software is Ciw [26]. Ciw is a discrete event simulation library written in Python that is fully documented and tested. The simulations constructed and studied in Sections **??** and 3 utilise this library and aid the overall reproducibility of this work.

1.1.3 Handling incomplete queue data

As is discussed in other parts of this section, the data available in this work is not as fine as in other comparative works. Without access to such distinct and detailed data — but with the aim of gaining insight from what is available — it is imperative that the gap left by the incomplete data be bridged.

Indeed, 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 where the line ends at incomplete queue data. [2] 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: [23, 38]. 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. [23] 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, [38] 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. [7, 12] 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. The dataset contains an administrative summary of 5,231 patients presenting COPD from February 2011 through March 2019 totalling 10,861 spells. A patient (hospital) spell is defined as the continuous stay of a patient using a hospital bed on premises controlled by a health care provider and is made up of one or more patient episodes [24].

The spells included in the dataset are described by 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 given 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.
- 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 [5, 30, 31].

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.

Due to a lack of information about the patients themselves — beyond their COPD-related admissions — the spells of the dataset were segmented using a variant of the k-means algorithm. 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 [16].

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 [29] 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 [34] 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. From this table, a number of helpful insights can be made. For instance, the needs of the spells in each cluster can be summarised succinctly:

- Cluster 0 represents those spells with relatively low clinical complexity but high resource requirements. The mean spell cost is almost four times the population average and the shortest spell is almost two weeks long. Moreover, the mean number of COPD-related admissions in the last year is higher for this cluster than any other indicating that the patients therein require more interactions with the system.

- Cluster 1 is the next largest segment and represents the spells with complex clinical profiles despite lower resource requirements. Specifically, the spells in this cluster have

		Cluster				Population
		0	1	2	3	
Characteristics	**Percentage of spells**	9.91	19.27	69.39	1.44	100.00
	Mean spell cost, £	8051.23	2309.63	1508.41	17888.43	2265.40
	Mean age	75.97	76.68	70.55	81.36	72.21
	COPD adm. last year	2.19	1.96	1.88	2.09	1.93
	Minimum LOS	12.82	-0.00	-0.02	48.82	-0.02
	Mean LOS	25.30	6.46	4.11	75.36	7.68
	Maximum LOS	51.36	30.86	16.94	224.93	224.93
	Median 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, %**	80.20	83.42	65.76	89.74	70.94
	PR, %	15.80	13.43	27.97	8.97	23.69
	SN, %	3.81	2.87	4.63	1.28	4.16
	Both, %	0.19	0.29	1.63	0.00	1.21
LTC prevalence	**Pulmonary disease, %**	100.00	100.00	100.00	100.00	100.00
	Diabetes, %	19.05	28.14	14.84	25.00	17.96
	AMI, %	13.85	22.93	8.76	16.03	12.10
	CHF, %	12.45	53.85	0.00	26.28	11.99
	Renal disease, %	7.53	19.54	1.92	17.95	6.10
	Cancer, %	7.62	12.23	2.93	10.90	5.30
	Dementia, %	6.88	21.26	0.00	26.92	5.17
	CVA, %	8.64	13.33	0.70	19.87	4.20
	PVD, %	4.37	7.69	2.27	5.77	3.57
	CTD, %	5.11	4.25	3.11	4.49	3.54
	Obesity, %	2.51	3.01	1.49	7.69	1.97
	Metastatic cancer, %	1.58	4.49	0.00	0.64	1.03
	Paraplegia, %	1.30	3.73	0.24	0.64	1.02
	Diabetic compl., %	0.19	0.86	0.48	1.92	0.54
	Peptic ulcer, %	1.58	0.81	0.23	1.28	0.49
	Sepsis, %	1.77	0.91	0.15	1.92	0.48
	Liver disease, %	0.28	0.48	0.23	0.00	0.28
	C. diff, %	0.74	0.10	0.01	0.64	0.11
	Severe liver disease, %	0.19	0.43	0.00	0.00	0.10
	MRSA, %	0.28	0.05	0.03	1.28	0.07
	HIV, %	0.00	0.00	0.03	0.00	0.02

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

the highest median CCI and number of LTCs, and the highest condition prevalences across all clusters but they have the second lowest length of stay and spell costs.

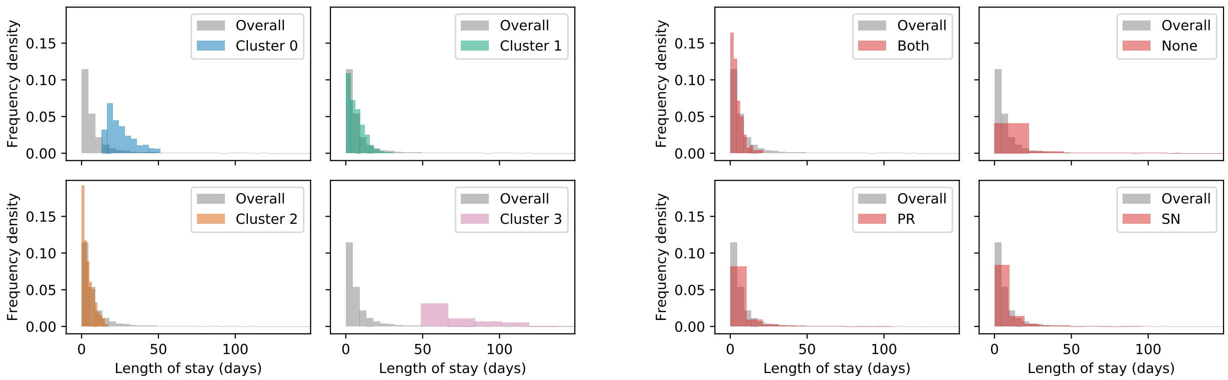
- Cluster 2 represents the majority of spells and those where resource requirements and clinical complexities are minimal; these spells have the shortest lengths, and the patients present with fewer diagnoses and a lower median CCI than any other cluster. In addition to this, the spells in Cluster 2 have the highest intervention prevalences and the lowest condition prevalences across all clusters.

- Cluster 3 represents the smallest section of the population but perhaps the most critical: spells with high complexity and high resource needs. The patients within Cluster 3 are the oldest in the population and are some of the most frequently returning despite having the lowest intervention rates. The lengths of stay vary between seven and 32 weeks, and the mean spell cost is almost eight times the population average. This cluster also has the second highest median CCI, and the highest median number of concurrent diagnoses.

The attributes listed in Table 1 can be studied beyond summaries such as these, however, and the figures remaining in this section aim to do that. In particular, the distributions for some of the clinical characteristics are shown in Figures 1 through 4 for each cluster. In addition to this, each of these figures also shows the distribution for the same attributes but by splitting the spell population by intervention rather than cluster. While this classical approach (i.e. of splitting a population based on a condition or treatment) can provide some insight into how the different interventions are used, it has been included to highlight the value added by segmenting the population on a number of attributes.

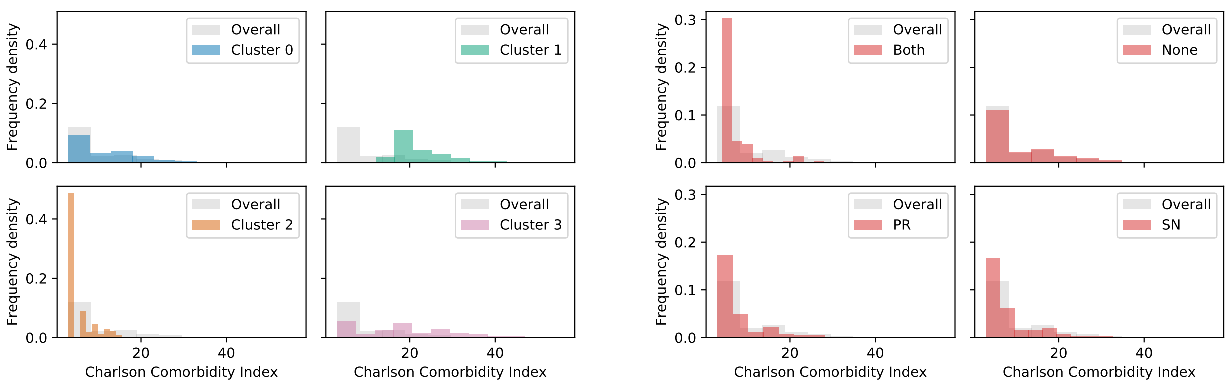
2 Estimating queuing parameters

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.



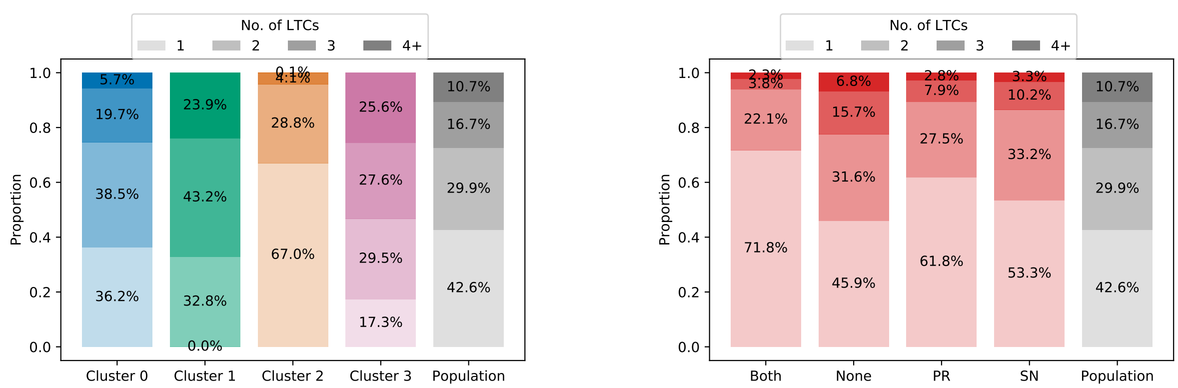
(a) (b)

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



(a) (b)

Figure 2: Histograms for CCI by (a) cluster and (b) intervention.



(a) (b)

Figure 3: Proportions of concurrent LTC counts presented by patients by (a) cluster and (b) intervention.

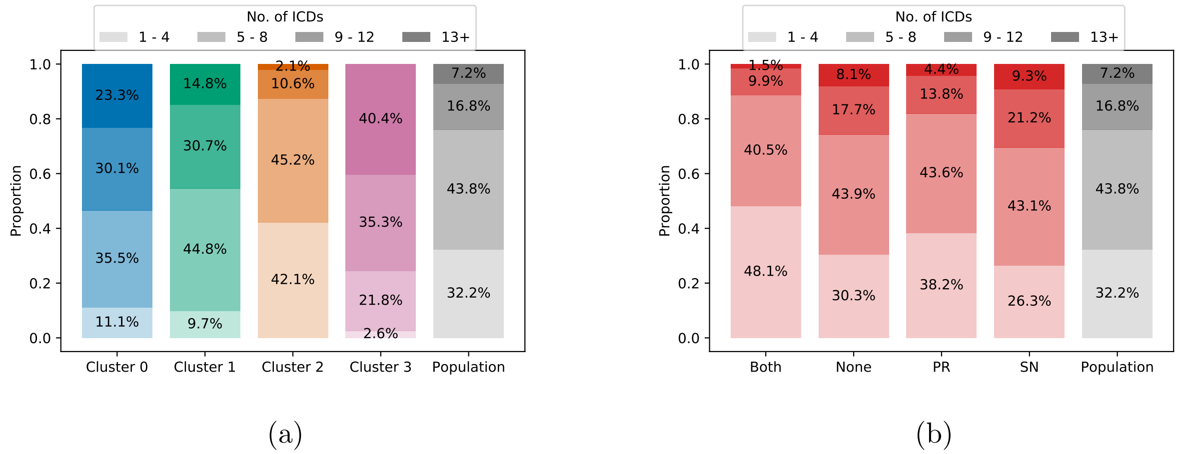


Figure 4: Histograms for number of ICDs by (a) cluster and (b) intervention.

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] 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.

[2] 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`.

[3] P. Bhattacharjee and P. K. Ray. Patient flow modelling and performance analysis of healthcare delivery processes in hospitals: A review and reflections. *Computers & Industrial Engineering*, 78:299 – 312, 2014. doi: 10.1016/j.cie.2014.04.016.

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

[5] P. F. Collins, R. J. Stratton, R. J. Kurukulaaratchy, and M. Elia. Influence of deprivation on health care use, health care costs, and mortality in copd. *International Journal of Chronic Obstructive Pulmonary Disease*, 13:1289–1296, 2018. doi: 10.2147/COPD.S157594.

[6] 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.

[7] 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.

[8] 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.

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

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

[11] B. G. Fitzpatrick. Issues in reproducible simulation research. *Bulletin of Mathematical Biology*, 81:1–6, 2019. doi: 10.1007/s11538-018-0496-1.

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

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

[14] 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.

[15] 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.

[16] 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.

[17] P. Ivie and D. Thain. Reproducibility in scientific computing. *ACM Computing Surveys*, 51(3), 2018. doi: 10.1145/3186266.

[18] A. Komashie, A. Mousavi, P. J. Clarkson, and T. Young. An integrated model of patient and staff satisfaction using queuing theory. *IEEE Journal of Translational Engineering in Health and Medicine*, 3:1–10, 2015. doi: 10.1109/JTEHM.2015.2400436.

[19] 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.

[20] F. B. Larsen, M. H. Pedersen, K. Friis, C. Glümer, and M. Lasgaard. 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), 2017. doi: 10.1371/journal.pone.0169426.

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

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

[23] 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.

[24] NHS Data Model and Dictionary. NHS Business Definitions: Hospital Provider Spell, 2020 (accessed 2020-06-15). URL `https://www.datadictionary.nhs.uk/data_dictionary/nhs_business_definitions/h/hospital_provider_spell_de.asp`.

[25] S. Olafsson, X. Li, and S. Wu. Operations research and data mining. *European Journal of Operational Research*, 87(3):1429 – 1448, 2008. doi: https://doi.org/10.1016/j.ejor.2006.09.023.

[26] 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.

[27] R. K. Palvannan and K. L. Teow. Queueing for healthcare. *Journal of Medical Systems*, 36:541–547, 2012. doi: 10.1007/s10916-010-9499-7.

[28] L. R. Pinto, F. C. C. de Campos, I. H. O. Perpétuo, and Y. C. N. M. B. Ribeiro. Analysis of hospital bed capacity via queuing theory and simulation. In *Proceedings of the Winter Simulation Conference 2014*, pages 1281–1292, 2014.

[29] 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.

[30] E. Sexton and D. Bedford. 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, 2016. doi: 10.1007/s11845-015-1359-5.

[31] M. C. Steiner, D. Lowe, K. Beckford, J. Blakey, C. E. Bolton, S. Elkin, W. D. C. Man, C. M. Roberts, L. Sewell, P. Walker, and S. J. Singh. Socioeconomic deprivation and the outcome of pulmonary rehabilitation in England and Wales. *Thorax*, 72(6):530–537, 2017. doi: 10.1136/thoraxjnl-2016-209376.

[32] 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.

[33] 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.

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

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] Álvaro Rebuge and D. R. Ferreira. Business process analysis in healthcare environments: A methodology based on process mining. *Information Systems*, 37(2):99 – 116, 2012. doi: https://doi.org/10.1016/j.is.2011.01.003.