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 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 [11] 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.

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, the handling of missing or incomplete data in healthcare settings and queuing theory applied to hospital systems.

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 [19, 23]. For instance, clinical profiling often forms part of the wider analysis where each segment can be summarised in a phrase or infographic [18, 22].

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 [3] 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 [21].

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 [9, 14]. This method has proved useful in the study of comorbidity patterns as in [1, 13] 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 [7]. 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 [15, 20]. In addition to k-means, hierarchical clustering methods can be effective if a suitable number of parts cannot be found initially [18]. Although, supervised hierarchical segmentation methods such as classification and regression trees (as in [10]) have been used where an existing, well-defined label is of particular significance.

1.1.2 Handling incomplete data

These are references for the estimation of service times (theoretically) [4, 8, 12].

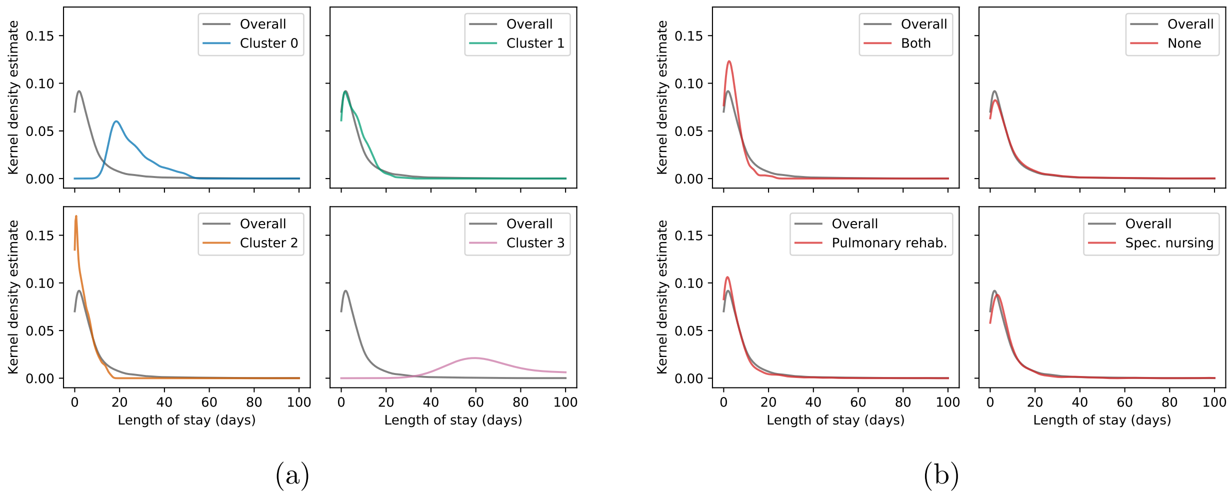


Figure 1: Kernel density estimate plots for length of stay by (a) cluster and (b) intervention

1.1.3 Queuing models

These are the principal queuing theory works by Erlang [5, 6]. Deadlock is an aspect of applied queuing theory of interest in recent literature [16]. The software used is Ciw [17].

1.2 Overview of the dataset

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 5,243 patients presenting COPD from February 2011 through March 2019 totalling 10,881 spells.

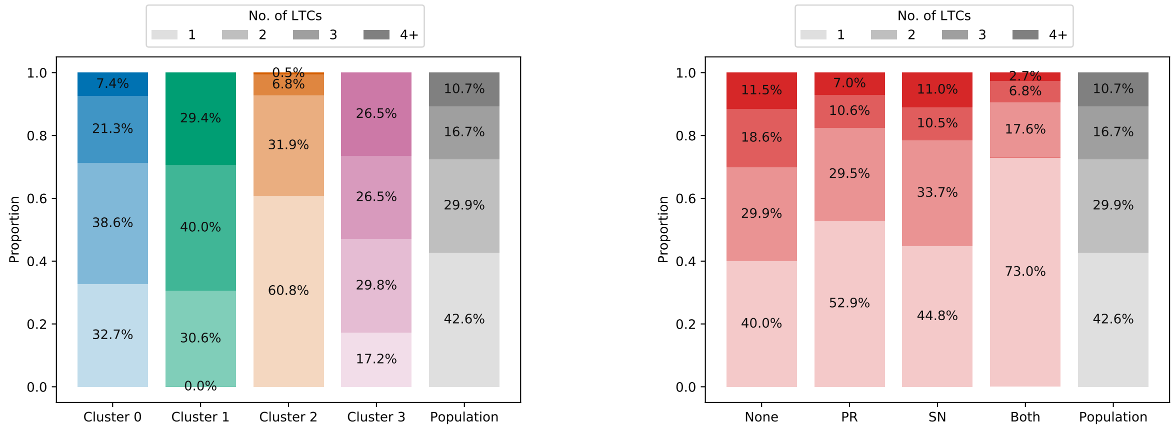
1.3 Cluster analysis

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

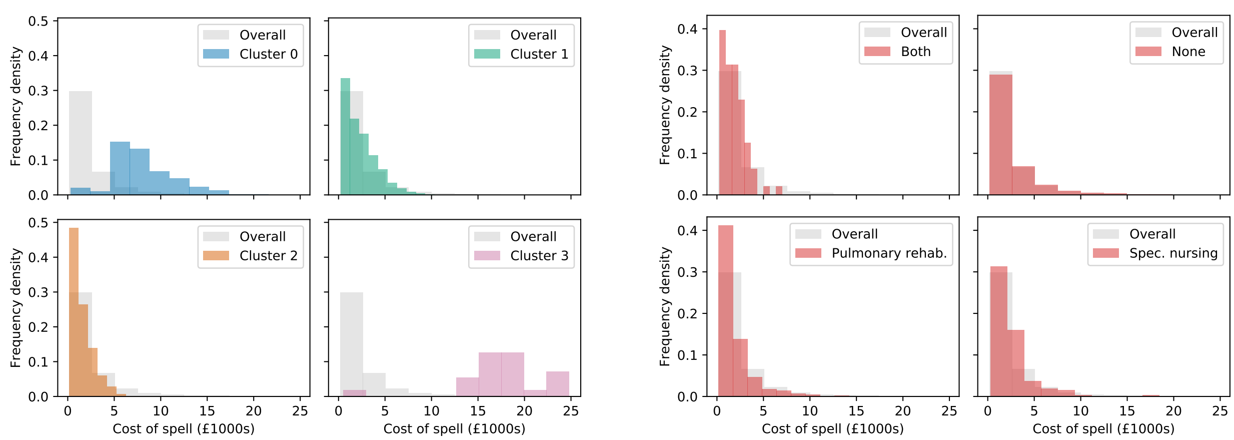
		Cluster				Population (mean)
		0	1	2	3	
Characteristics	**Mean spell cost**	8083.69	2312.39	1509.22	17847.80	2280.54
	Mean age	76.15	77.16	71.01	81.36	72.22
	Mean COPD adm. in year	1.91	1.51	1.31	1.98	1.29
	Min. LOS	12.82	-0.00	-0.02	48.82	5.41
	Mean LOS	25.30	6.45	3.79	74.65	7.47
	Max. LOS	51.36	30.86	16.94	224.93	10.40
	Median no. of LTCs	2.00	3.00	1.00	3.00	2.00
	Median no. ICDs	9.00	8.00	5.00	11.00	6.58
	Median CCI	9.00	20.00	4.00	18.00	9.72
Intervention prevalence	**None, %**	80.26	83.40	65.76	89.81	70.95
	Pulmonary rehab., %	15.77	13.41	27.96	8.92	23.66
	Spec. nursing, %	3.78	2.91	4.63	1.27	4.16
	Both, %	0.18	0.29	1.66	0.00	1.22
LTC prevalence	**Pulmonary disease, %**	100.00	100.00	100.00	100.00	100.00
	Diabetes, %	20.12	28.70	16.71	25.83	20.07
	CHF, %	14.47	54.62	0.00	27.15	14.83
	AMI, %	14.35	23.42	10.86	16.56	14.28
	Renal disease, %	8.59	22.52	3.02	18.54	8.50
	Cancer, %	8.12	13.69	4.11	10.60	6.94
	Dementia, %	7.53	20.08	0.00	25.17	6.10
	CVA, %	9.65	15.01	1.21	19.21	5.85
	PVD, %	5.06	8.90	3.18	5.96	4.78
	CTD, %	5.18	5.07	4.01	4.64	4.42
	Obesity, %	2.82	3.75	2.22	7.95	2.78
	Metastatic cancer, %	1.88	5.70	0.00	0.66	1.54
	Paraplegia, %	1.29	3.96	0.25	0.66	1.23
	Sepsis, %	2.12	1.25	0.25	1.99	0.76
	Peptic ulcer, %	1.76	1.04	0.35	1.32	0.72
	Diabetic compl., %	0.24	0.69	0.33	1.99	0.44
	Liver disease, %	0.35	0.56	0.33	0.00	0.37
	Severe liver disease, %	0.24	0.63	0.00	0.00	0.17
	C. diff, %	0.82	0.14	0.03	0.66	0.17
	MRSA, %	0.35	0.07	0.05	1.32	0.12
	HIV, %	0.00	0.00	0.03	0.00	0.02

Table 1: A summary of patient-level clinical attributes and disease prevalence by cluster and by population



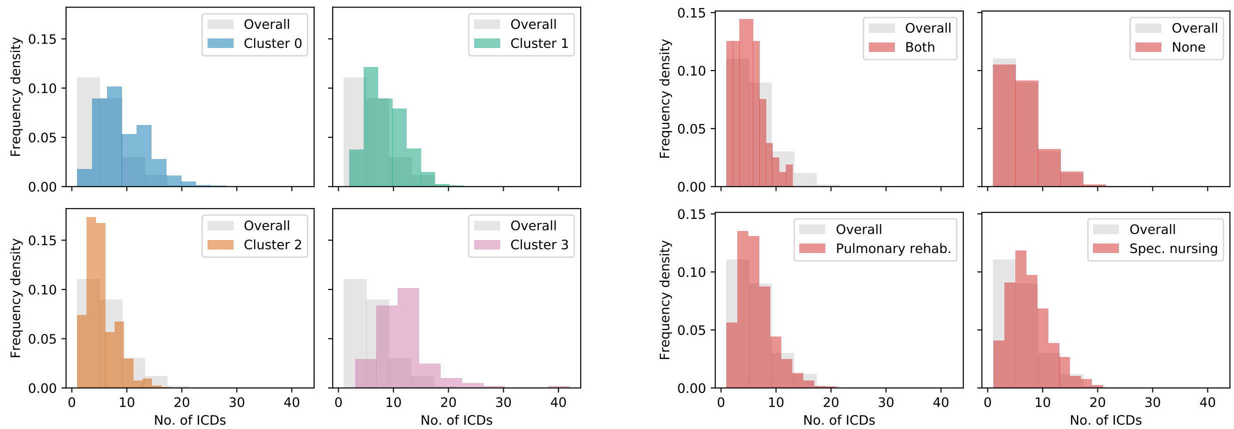
(a) (b)

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



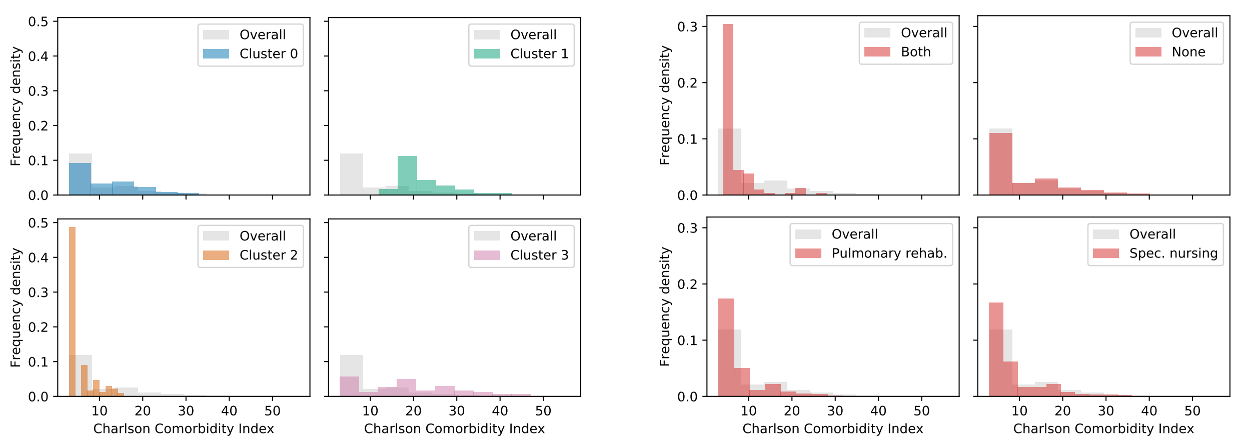
(a) (b)

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

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] 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.

[4] 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.

[5] 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.

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

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

[8] 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.

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

[10] 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.

[11] 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.

[12] D. Kumar, A. Tantawi, and L. Zhang. Estimating model parameters of adaptive software systems in real-time. In D. Ardagna and L. Zhang, editors, *Run-time Models for Self-managing Systems and Applications*, chapter 3, pages 45–71. 2010. doi: 10.1007/978-3-0346-0433-8_3.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.