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.

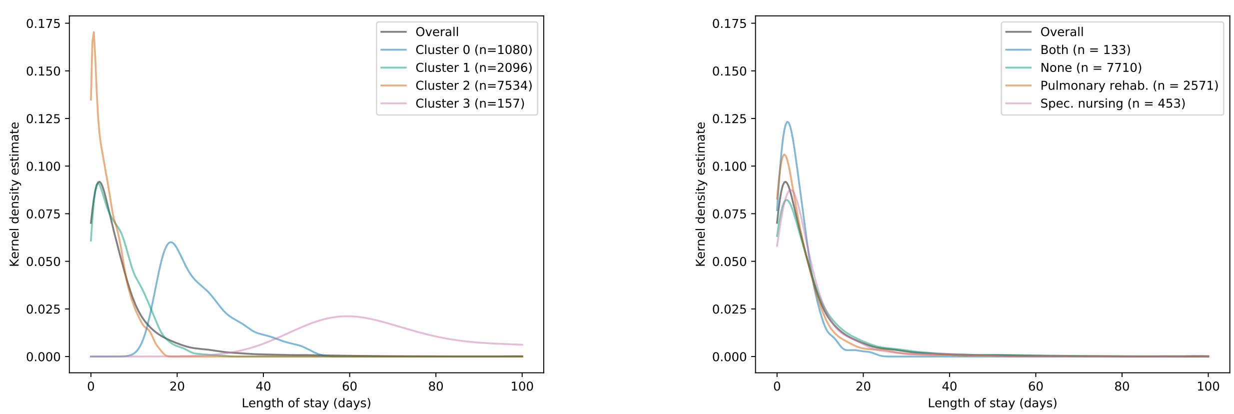
1 Introduction

Population health research is becoming increasingly based on data-driven methods (as opposed to those designed by clinical experts) through the advent of accessible software and an abundance of electronic data. To better understand the needs and behaviours of a population, it can be helpful 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 analysis is of patient flow and queuing systems.

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 meso-level population clustering and a multi-class queuing model.

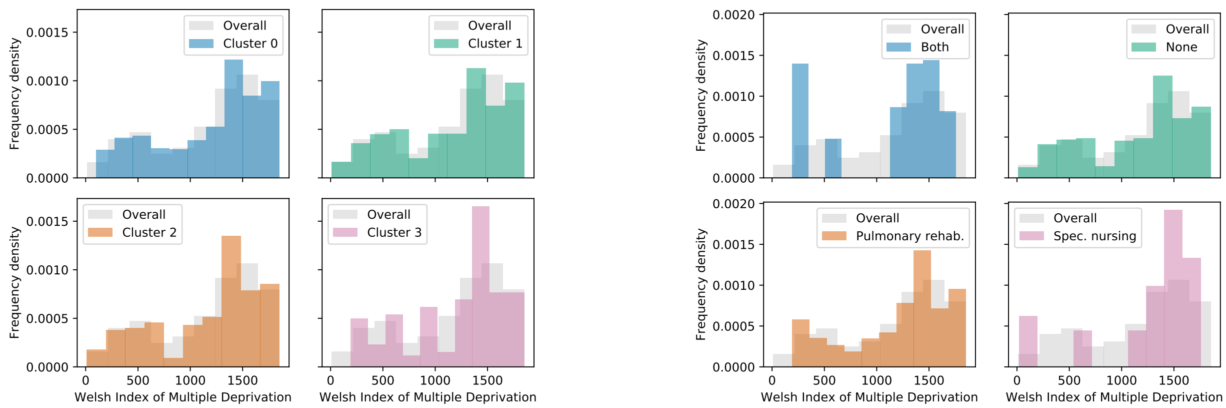
		Cluster 0	1	2	3	Population (mean)
Characteristics	**COPD admissions last year**	1.91	1.51	1.31	1.98	1.29
	Min. LOS	12.82	0.01	-0.02	48.82	5.41
	Mean LOS	25.33	6.46	3.79	74.65	7.48
	Max. LOS	51.36	30.86	16.94	224.93	10.40
	No. of LTCs	2.05	3.10	1.47	2.70	2.00
	Median no. ICDs	9.00	8.00	5.00	11.00	6.58
	Median CCI	9.00	20.00	4.00	18.00	9.72
LTC Prevalence	**Acute myocardial infection, %**	14.27	23.42	10.84	16.56	14.26
	Cerebrovascular accident, %	9.67	15.01	1.19	19.21	5.84
	Congestive heart failure, %	14.50	54.55	0.00	27.15	14.82
	Connective tissue disorder, %	5.19	5.07	4.01	4.64	4.42
	Dementia, %	7.55	20.08	0.00	25.17	6.11
	Diabetes, %	20.17	28.70	16.72	25.83	20.08
	Liver disease, %	0.35	0.56	0.33	0.00	0.37
	Peptic ulcer, %	1.77	1.04	0.38	1.32	0.73
	Peripheral vascualar disease, %	5.07	8.90	3.18	5.96	4.78
	Pulmonary disease, %	100.00	100.00	100.00	100.00	100.00
	Cancer, %	8.02	13.76	4.11	10.60	6.95
	Diabetic complications, %	0.24	0.69	0.33	1.99	0.44
	Paraplegia, %	1.30	3.96	0.25	0.66	1.23
	Renal disease, %	8.61	22.52	3.03	18.54	8.51
	Metastatic cancer, %	1.77	5.77	0.00	0.66	1.55
	Severe liver disease, %	0.24	0.63	0.00	0.00	0.17
	HIV, %	0.00	0.00	0.03	0.00	0.02
	C. diff, %	0.83	0.14	0.03	0.66	0.17
	MRSA, %	0.35	0.07	0.05	1.32	0.12
	Obesity, %	2.83	3.75	2.22	7.95	2.78
	Sepsis, %	2.12	1.25	0.25	1.99	0.77

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



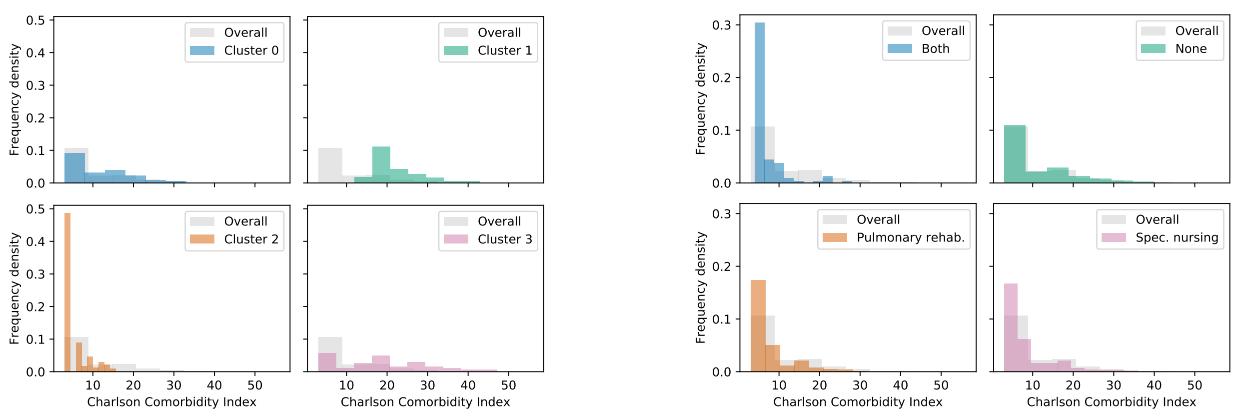
(a) (b)

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



(a) (b)

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



(a) (b)

Figure 3: Histograms for CCI by (a) cluster and (b) intervention

1.1 Literature review

1.2 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 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).