Informatics and Technology in Clinical Care and Public Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210851

Using Machine Learning to Improve Personalised Prediction: A Data-Driven Approach to Segment and Stratify Populations for Healthcare

Will YUILL[a,b,1] and Holger KUNZ[a,2]
[a] *Institute of Health Informatics, University College London, UK*
[b] *Hertfordshire County Council, UK*

Abstract. Population Health Management typically relies on subjective decisions to segment and stratify populations. This study combines unsupervised clustering for segmentation and supervised classification, personalised to clusters, for stratification. An increase in cluster homogeneity, sensitivity and positive predictive value was observed compared to an unlinked approach. This analysis demonstrates the potential for a cluster-then-predict methodology to improve and personalise decisions in healthcare systems.

Keywords. Population Health Management, segmentation, stratification, clustering, machine learning

1. Introduction

Population Health Management (PHM) is increasingly being adopted in England to improve outcomes for individuals by personalising services to address their health and care needs in a way that recognises that health is determined more by socio-economic factors than by healthcare provision itself [1]. However, personalisation of interventions on an individual level is not feasible because of resource constraints [2]. Instead, PHM proposes designing systems around defined segments of the population, further targeting individuals through stratification by risk of an adverse event, such as hospital readmission or onset of disease. Segmentation and stratification can improve patient care and management and inform the design of care systems [3,4]. The foundations of segmentation and stratification in healthcare have already been defined by Garfield [5], but the increasing availability of data provides opportunity for new methods to be used.

There are two broad approaches to segmentation [6]. The traditional approach is to use *a priori* groups based on expert knowledge, for instance segmenting patients by morbidity, age, or disease. This approach can have limitations:

- The number of features by which to segment must remain small because the number of groups can expand rapidly, even exponentially in some cases.

[1] Corresponding Author, Will Yuill, Institute of Health Informatics, University College London, London, UK, & Hertfordshire County Council, UK, E-mail: will.yuill@hertfordshire.gov.uk

[2] Corresponding Author, Holger Kunz, Institute of Health Informatics, University College London, London, UK, E-mail: h.kunz@ucl.ac.uk

- With few features, important differentiators between people are missed.
- Variables such as age require arbitrary breakpoints to be defined.
- With few segments and features, entropy is likely to be high.

The second, more data-driven approach has been adopted recently and addresses many of the limitations of *a priori* methods, allowing more variables to be used, and for feature breakpoints to be derived naturally [6,7]. However, this approach requires a large amount of data and may result in less readily comprehensible clusters. For example, while *a priori* segmentation allows precise classification of segments [2], data-driven methods can result in more nuanced groupings, requiring interpretation [6]. Therefore, the purpose of the segmentation is likely to influence the method chosen.

Segmentation can often group patients of similar types but differing magnitudes of need. This is where stratification can assist. Typically, stratification is performed with ordinary regression models (developed on a wholly different population) and applied equally to every segment. However, by using segmentation, we postulate there are meaningful differences between groups and therefore predictors of risk may differ. Not accounting for this could reduce the performance of stratification models [8,9].

This study therefore explored the potential for data-driven segmentation, coupled with stratification personalised to each segment, to achieve better performance than existing models when predicting the risk of emergency readmission within 30 days of discharge. This cluster-then-predict method has been effective in other settings [10].

2. Methodology

The study population comprised 78,786 admission episodes in the NHS Secondary Uses Service dataset for patients registered with a single Clinical Commissioning Group in England in the fiscal year 2020–2021. For each admission the following data was extracted with exclusions applied in accordance with Billings et al. [4]:

- Person: age, deprivation assessed via Index of Multiple Deprivation 2019 [11].
- Health: morbidities in inpatient and outpatient records in the three years prior to admission classified in the Charlson Comorbidity Index [12].
- Care: NHS Provider Trust, count of emergency admissions in the year prior, emergency admission in the past 30 days, whether the current admission was an emergency, emergency readmission within 30 days (target variable).

Three methods were implemented using the methodology summarised in Figure 1.

2.1. Model 1: a priori Segmentation and Traditional Stratification

Morbidities were assigned a score [12] and these were summed to split episodes into groups ‘0’, ‘1-2’, ‘3-4’, ‘5+’. The PARR-30 algorithm [4] was applied to calculate readmission risk. The cost of a false negative was set at three times the cost of a false positive and used to select a threshold to predict readmission [13]. Homogeneity of clusters was assessed through Silhouette scores performed on a 30,000-record subset.

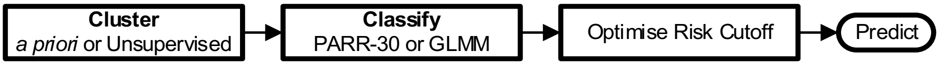

Figure 1. Flowchart of methods used to cluster-then-predict using both supervised and unsupervised learning.

2.2. Model 2: Traditional Segmentation and Personalised Stratification

Patients were segmented as in Model 1, but a generalised logistic mixed model (GLMM) defined in Eq. (1) was used to calculate readmission risk with the threshold optimised as in Model 1. Rather than fitting regressions to each segment, GLMMs allow for pooling between segments, mitigating small groups, and allowing greater scaling. Intercept and slope were allowed to vary by segment. Only feature used in Model 1 were included to facilitate comparison. Providers with fewer than 1,000 instances were recoded as ‘other’.

$$\begin{gathered} Readmission = f(Age, CurrentEmergAdmi, EmergAdmiLast30d, EmergAdmiLast1yr, \\ CHF, PVD, CPD, ChronicDM, Renal, SolidCancer, OtherCancer, MildLiver, Dementia, \\ ModOrSevereLiver, HemiParaPlegia, ProviderTrust, IndexMultipleDeprivationScore) \end{gathered} \tag{1}$$

2.3. Model 3: Data-Driven Segmentation and Personalised Stratification

To investigate whether discovering natural clusters in the data would result in improved predictions, unsupervised clustering was undertaken using k-prototypes in order for both binary and continuous variables to be used [14]. Features were limited to those in PARR-30, excluding NHS Provider Trust because of its high cardinality. Continuous features were scaled and centered. The number of clusters was set to maximise Silhouette score and a GLMM as defined in Model 2 was fitted to the result.

3. Results

Table 1 summarises the performance for each model. Performance was similar across accuracy, area under the curve (AUC) and specificity. Small but significant improvements were seen in positive predictive value (PPV) when comparing Models 1 and 2, with a further significant improvement in sensitivity when comparing Models 1 and 3. Performance between segments varied, for instance PPV in Model 3 ranged from 0.19–0.44. The clustering of Models 1 and 2 was poor with a mean Silhouette score of -0.12 compared to 0.18 in Model 3, which found five clusters to be optimum.

Table 1. The performance of each model with 95% confidence intervals in brackets.

Model	Accuracy	AUC	Sensitivity	Specificity	PPV
1	0.88 (0.88-0.88)	0.73 (0.72-0.74)	0.20 (0.19-0.21)	0.95 (0.95-0.95)	0.30 (0.29-0.32)
2	0.89 (0.88-0.89)	0.75 (0.74-0.76)	0.22 (0.21-0.23)	0.96 (0.95-0.96)	0.34 (0.33-0.35)
3	0.88 (0.88-0.89)	0.75 (0.75-0.76)	0.25 (0.24-0.26)	0.95 (0.95-0.95)	0.34 (0.34-0.35)

4. Discussion

This study implemented three models to improve PHM segmentation and stratification through a cluster-then-predict methodology. Personalising risk stratification as in Model 2 resulted in small but significant improvements to predictive performance over Model 1, suggesting that integrating segmentation and stratification approaches can improve understanding of patient risk by personalising prediction to segments. The variance between the performance of different segments remained, suggesting that the data used may not provide sufficient information to accurately predict risk or that segments were not sufficiently homogenous, as indicated by Silhouette score. The creation of more homogenous clusters in Model 3 provided further small but significant improvements to stratification, suggesting that increases in homogeneity result in better predictions, even if these increases are small and homogeneity remains poor.

These results suggest that, with further optimisation, both more homogenous clustering methods and personalisation of stratification can provide more accurate risk prediction than traditional techniques. Given the similarity in performance between Models 2 and 3, a sufficiently homogenous *a priori* segmentation could provide both well-defined segments and improved risk prediction and, where the understandability of segments is important, this may be preferred. There is clear evidence that the addition of primary care and other datasets can result in both more homogenous segments (be this *a priori* or data-driven) and accurate classification [3,7], and future work could include this data where feasible. Further improvements may also be possible with methods that bind segmentation and stratification more closely together, such as ToPs/R [15]. The development of these methods offers to inform the development care that better meet the diverse needs of the population. Personalised medicine could be supported with the identification of appropriate clusters and to address the specific needs of each group.

References

[1] Hood CM, Gennuso KP, Swain GR, Catlin BB. County health rankings: Relationships between determinant factors and health outcomes. Am J Prev Med. 2016 Feb 1;50(2):129–35.

[2] Lynn J, Straube BM, Bell KM, Jencks SF, Kambic RT. Using population segmentation to provide better health care for all: The “Bridges to Health” model. Milbank Q. 2007 Jun;85(2):185–208.

[3] Vuik SI, Mayer E, Darzi A. Enhancing risk stratification for use in integrated care: a cluster analysis of high-risk patients in a retrospective cohort study. BMJ Open. 2016 Dec 1;6(12):e012903.

[4] Billings J, Blunt I, Steventon A, Georghiou T, Lewis G, Bardsley M. Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30). BMJ Open. 2012 Aug 10;2(4):e001667.

[5] Garfield SR. The delivery of medical care. Sci Am. 1970 Apr;222(4):15–23.

[6] Nnoaham KE, Cann KF. Can cluster analyses of linked healthcare data identify unique population segments in a general practice-registered population? BMC Public Health. 2020 May 27;20(798):1–10.

[7] Vuik S. On the application of data-driven population segmentation to design patient-centred integrated care. Imperial College London; 2017. 189 p.

[8] Li Y, Sperrin M, Belmonte M, Pate A, Ashcroft DM, van Staa TP. Do population-level risk prediction models that use routinely collected health data reliably predict individual risks? Sci Rep. 2019 Aug 2;9(1):11222.

[9] Xia E, Du X, Mei J, Sun W, Tong S, Kang Z et al. Outcome-driven clustering of acute coronary syndrome patients using multi-task neural network with attention. In: Studies in Health Technology and Informatics. Amsterdam: IOS Press; 2019. p. 457–61.

[10] Soni R, Mathai KJ. Improved Twitter Sentiment Prediction through Cluster-then-Predict Model. Int J Comput Sci Netw. 2015 Aug;4(4):559–63.

[11] Ministry of Housing, Communities & Local Government. English indices of deprivation [Internet]. 2019. Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 [Accessed 01 Sep 2021].

[12] Gasparini A. comorbidity: An R package for computing comorbidity scores. J Open Source Software. 2018;3(23):648.

[13] Billings J, Dixon J, Mijanovich T, Wennberg D. Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. BMJ. 2006 Aug 12;333(7563):327–30.

[14] Szepannek G. clustMixType: User-friendly clustering of mixed-type data in R. R J. 2019;10(2):200–8.

[15] Yoon J, Zame WR, Banerjee A, Cadeiras M, Alaa AM, van der Schaar M. Personalized survival predictions via Trees of Predictors: An application to cardiac transplantation. PLoS One. 2018 Mar 28;13(3):e0194985.