Enhancing risk stratification for use in integrated care - A cluster analysis of high-risk patients

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2016-012903
Article Type:	Research
Date Submitted by the Author:	02-Jun-2016
Complete List of Authors:	Vuik, Sabine; Imperial College London, Institute of Global Health Innovation Mayer, Erik; Imperial College London, Dept. of Biosurgery and Surgical Technology Darzi, Ara; Imperial College London, Institute of Global Health Innovation
Primary Subject Heading:	Health services research
Secondary Subject Heading:	Research methods, Patient-centred medicine, Evidence based practice, General practice / Family practice
Keywords:	Risk management < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Organisation of health services < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, STATISTICS & RESEARCH METHODS

Enhancing risk stratification for use in integrated care - A cluster analysis of high-risk patients

Sabine I Vuik, Erik Mayer, Ara Darzi

Sabine I Vuik, Policy Fellow, Institute of Global Health Innovation

Imperial College, 10th Floor, St Mary's Hospital, Praed Street, London, W1 2NY, UK

Erik Mayer, Clinical Senior Lecturer, Department of Surgery

Imperial College, 10th Floor, St Mary's Hospital, Praed Street, London, W1 2NY, UK

Ara Darzi, Professor in Surgery, Department of Surgery

Imperial College, 10th Floor, St Mary's Hospital, Praed Street, London, W1 2NY, UK

Correspondence to: Sabine Vuik s.vuik@imperial.ac.uk, +44(0) 795 714 0479

Keywords: risk prediction, integrated care, care utilisation, emergency hospitalisation, high-risk patients

Word count (excl. title page, abstract, references, tables and figures): 2,456

ABSTRACT

Objective: To show how segmentation can enhance risk stratification tools for integrated care, by providing insight into different care utilisation patterns within the high-risk population.

Design: A retrospective cohort study. A risk score was calculated for each person using a logistic regression, which was then used to select the top 5% high-risk individuals. This population was segmented based on utilisation of different care settings using a k-means cluster analysis. Data from 2008 to 2011 was used to create the risk score and segments, while 2012 data was used to understand the predictive abilities of the models.

Setting and participants: Data was collected on primary care use (CPRD) and secondary care use (HES) for a random sample of 300,000 English patients.

Main measures: The high-risk population was segmented based on utilisation of four different care settings: emergency acute care, elective acute care, outpatient care and GP care

Results: While the risk strata predicted care utilisation at a high level, within the high-risk population utilisation varied significantly. Four different groups of high-risk patients could be identified. These four segments had distinct utilisation patterns across care settings, reflecting different levels and types of care needs. The 2008-2011 utilisation patterns of the four segments were consistent with the 2012 patterns.

Discussion: Cluster analyses revealed that the high-risk population is not homogeneous, as there exist four groups of patients with different needs across the care continuum. Since the patterns were predictive of future care use, they can be used to develop integrated care programmes tailored to these different groups.

Conclusions: Utilisation-based segmentation augments risk stratification by identifying patient groups with different care needs, around which integrated care programmes can be designed.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- This study uses patient-level linked primary and secondary care administrative data
- Rather than focusing only on emergency care, this study looks at patterns of utilisation across different care settings to support the development of integrated care programmes

- Where previous studies have focused on how to identify or manage high-risk patients, this study explores the different patient groups within the high-risk stratum
- The data used was for a random sample of English patients, and may not reflect local trends
- No data was available in linked format for other care settings, such as A&E, mental health, community and social care

BACKGROUND

In healthcare, a small number of patients accounts for a disproportionally large share of utilisation.[1 2] Identifying and targeting this group can be done through risk stratification. Risk stratification divides a population based on different levels of risk of a specific outcome, and is a core process to achieve integrated, personalised care.[3-5] For each stratum, a tailored care model can be developed which addresses the specific needs of the patients. Many of the interventions for high-risk patients are primary care-led integrated care programmes, like virtual wards, case management, and enhanced services and access. [4 6-11]

Risk stratification methods often focus on predicting emergency hospitalisations.[3 12-15] Unplanned hospitalisations, including readmissions, are chosen because they are costly for a health system, may indicate low quality care, and have a negative impact on patient experience.[16 17] As such, unplanned hospitalisations are reflective of all elements of the triple aim of healthcare – quality of care, patient experience and cost[18] – and can be considered a 'triple fail event'.[16] Moreover, since preventing emergency hospitalisations to the acute setting requires effective primary care, they are also an important metric for integrated care.[19]

However, risk stratification based on emergency hospitalisations has important limitations. Firstly, this approach only looks at one element of care. While the risk of an emergency hospitalisation can be expected to correlate with overall use of emergency acute care, utilisation of other care services may vary. A patient with an emergency hospitalisation may be under treatment with a specialist; or regularly visit a general practitioner (GP); or not access ambulatory care at all. In order to design effective integrated care programmes that link up the appropriate care providers, understanding care use across all settings is crucial.

Secondly, detailed information on the characteristics of the high-risk patients, such as age, morbidities and socio-economic status, is lost in the final risk score. All patients who end up in the top stratum have high risk scores, but the factors driving this high score can be very different. When developing interventions, these should be taken into account to understand which patients are most likely to respond to different interventions.[12 20]

The aim of this study is to show how utilisation-based segmentation can enhance risk stratification tools used for integrated care by, firstly, taking into account care utilisation across multiple care settings and, secondly, providing insight into the characteristics of different patient groups within the high-risk stratum.

METHODS

Study design

To show how segmentation can augment risk stratification, we applied both methods to a large patient database. We first trained a risk prediction model to generate risk scores for each patient. Based on these risk scores, we identified the high-risk patient population. In this group we applied a cluster analysis to a range of different utilisation variables. The different clusters were analysed and profiled to understand the different patient types that exist within a high-risk group.

The analyses were conducted for hypothetical "historic" (2008-2011) and "future" (2012) datasets. The historic dataset reflects the information that would be available to healthcare professionals conducting risk stratification and cluster analysis at the end of 2011, while the future dataset was used to understand how accurately the models predicted actual utilisation in the following year.

Data

A dataset covering primary and secondary care use for a random sample of 300,000 English patients was constructed from Clinical Practice Research Datalink (CPRD) and Hospital Episode Statistics (HES) data (CPRD ISAC approval under protocol 14_211R).

Patients were eligible for inclusion if they were registered with a CPRD-participating GP practice during the entire study period of 2008 up to and including 2012, and if their HES records could be linked to CPRD. In England, Clinical Commissioning Groups (CCG) are responsible for the planning and commissioning of care for local populations. The sample size in this study was set at 300,000, which is similar to the population of a CCG in the 75th percentile,[21] to reflect a typical local population in England.

The final dataset included patient demographics, long-term condition (LTC) diagnoses and utilisation variables. We selected four high-level utilisation variables for the cluster analysis of high-risk patients: inpatient emergency hospitalisations, inpatient nonemergency hospitalisations, outpatient attendances and GP visits. These utilisation variables were used to reflect different care settings that may be incorporated in integrated care models.

Risk stratification

We calculated our own risk prediction score, reflecting predictor variables used in PARR, the Combined Predictive Model and other commonly used risk prediction algorithms. The risk model was trained to predict emergency hospitalisations in 2012, using a stepwise logistic regression.[14 22] The number of emergency hospitalisations in 2011 was included as one of the predictor variables, as well as a range of other variables detailed in appendix 1.

The logistic regression on the training set excluded a number of diagnosis variables after step-wise elimination, as well as the over 75+ flag.

To validate the model, a split sample validation method was used. Using the random sample function of SPSS,[23] half of the sample was defined as the training set and the other half as the test set. Applying the risk model to the test set, the area under the Receiver Operator Curve (ROC) was 0.75. This is in line with other models predicting emergency hospitalisations, which range from 0.55 to 0.83.[13 36] The test population was stratified into three groups, which are comprised of the top 5% highest risk patient ("High risk"), the top 5-20% ("Medium risk") and the remaining 80% of the population ("Low risk"), in accordance with general risk stratification practice.[2 15 17]

Segmentation

For the segmentation analysis the k-means algorithm was used to cluster the patients based on their historic utilisation. This method was selected as it is efficient and produces roughly similar sized segments.[24] Clustering solutions ranging from 2 to 8 clusters were explored for the high-risk stratum. To identify the optimal number of clusters, the Pseudo-F statistic was calculated for all the clustering solutions using STATA.[25] This statistic is commonly used in healthcare clustering studies,[26-30] and is one of the best criteria to determine the number of clusters.[31] It compares the between-cluster to the within-cluster sum-of-squares, and a large Pseudo-F statistic indicates distinct clusters.[32]

Analysis

To create profiles for the segments, the utilisation variables as well as demographic characteristics were analysed to see if they differed significantly across segments. For the non-Normal utilisation and LTCs count variables, a Kruskal-Wallis test was used. For the continuous age and risk score variables an ANOVA test was used, and for the binary morbidity variables and the 2012 emergency hospitalisation flag a Chi square test. Where these tests found significant variation across segments, the results were then explored pair-wise between segments to identify which segment or segments were significantly different from others. For this, Mann-Whitney U tests, Student t-tests, and z-tests were used, respectively. To account for the multiplicity problem that occurs when performing multiple tests, the Bonferroni method was used to adjust the significance level.[33-35]

RESULTS

The final dataset contained 298,111 people with a complete record across the variables, of which 149,320 observations ended up in the test set used for the analyses below. When the population was stratified based on risk, predictive variables such as age, long-term conditions and historic care utilisation were all found to increase with each risk stratum (see table 1). In addition to historic utilisation, future utilisation of all care types also increased consistently with the risk strata.

Table 1: Strata characteristics

	High risk	*Medium risk*	*Low risk*	***Total population***
Number of people	7,466	22,398	119,456	**149,320**
Predicted proportion with any emergency hospitalisations in 2012	27%	9%	3%	**5%**
Actual proportion with any emergency hospitalisations in 2012	27%	11%	3%	**5%**
Age at end of study period, mean	75	65	40	**45**
Number of long-term conditions, mean	1.7	0.7	0.1	**0.3**
Number of emergency hospitalisations per year (historic), mean	0.5	0.1	0.0	**0.1**
Number of nonemergency hospitalisations per year (historic), mean	0.6	0.3	0.1	**0.1**
Number of outpatient attendances per year (historic), mean	5.8	3.0	0.8	**1.4**
Number of GP visits per year (historic), mean	15.7	9.6	3.4	**5.0**
Number of emergency hospitalisations per year (future), mean	0.4	0.1	0.0	**0.1**
Number of nonemergency hospitalisations per year (future), mean	0.5	0.4	0.1	**0.2**
Number of outpatient attendances per year (future), mean	6.1	3.4	1.0	**1.6**
Number of GP visits per year (future), mean	17.0	10.5	3.8	**5.5**

For the high-risk population, k-means cluster analyses were performed for 2- to 8-clusters and the pseudo-F statistics was obtained for each solution. A peak was observed around the 3- and 4-cluster solutions. Exploring these two sets of clusters, the 4-cluster solution included an additional, contrasting utilisation pattern and was therefore selected.

The cluster analysis aims to optimise the distance between groups for the clustering variables, and statistical tests confirm that historic utilisation is significantly different across segments (see table 2) In addition, non-clustering variables, including future utilisation, age, number of long-term conditions and most disease prevalence variables, also differ significantly across the clusters.

Table 2: Clusters within the high-risk population

	Cluster				ANOVA/ Kruskal-Wallis/ Chi square test
	1	2	3	4	
Number of people	1967	1807	1831	1861	
Predicted proportion with any emergency hospitalisations in 2012 (based on average risk score), %	21 ***	38 ***	20 ***	31 ***	**AN: <0.000**
Actual proportion with any emergency hospitalisations in 2012, %	19 **	35 **	21 **	34 **	**Chi: <0.000**
Age at end of study period, mean	79 ***	67 ***	83 ***	71 ***	**AN: <0.000**
Number of long-term conditions, mean	1.8 **	2.0 **	1.4 ***	1.7 ***	**KW: <0.000**
Number of emergency hospitalisations per year (historic), mean	0.1 **	0.9 ***	0.2 **	0.8 ***	**KW: <0.000**
Number of nonemergency hospitalisations per year (historic), mean	1.0 ***	1.1 ***	0.1 ***	0.1 ***	**KW: <0.000**
Number of outpatient attendances per year (historic), mean	7.9 ***	9.3 ***	2.5 ***	3.3 ***	**KW: <0.000**
Number of GP visits per year (historic), mean	17.6 ***	16.7 ***	15.9 ***	12.5 ***	**KW: <0.000**
Number of emergency hospitalisations per year (future), mean	0.3 **	0.6 **	0.3 **	0.6 **	**KW: <0.000**
Number of nonemergency hospitalisations per year (future), mean	0.7 **	0.9 **	0.3 ***	0.3 ***	**KW: <0.000**
Number of outpatient attendances per year (future), mean	7.7 ***	9.1 ***	3.4 ***	4.2 ***	**KW: <0.000**
Number of GP visits per year (future), mean	18.5 ***	17.9 **	17.5 **	14.2 ***	**KW: <0.000**
Prevalence of AMI, %	15 ***	23 ***	10 ***	19 ***	**Chi: <0.000**
Prevalence of asthma, %	28 *	26	24 *	25	**Chi: 0.028**
Prevalence of cancer, %	26 ***	22 ***	8 ***	5 ***	**Chi: <0.000**
Prevalence of cerebrovascular disease, %	9 **	15 **	10 **	18 **	**Chi: <0.000**

Prevalence of congestive heart failure, %	8 ***	13 **	5 ***	13 **	**Chi: <0.000**
Prevalence of COPD, %	18 *	17 *	13 ***	18 *	**Chi: <0.000**
Prevalence of dementia, %	3 **	3 **	5 **	7 **	**Chi: <0.000**
Prevalence of diabetes, %	28 **	22 **	28 **	22 **	**Chi: <0.000**
Prevalence of HIV/AIDS, %	0	0	0	0	Chi: 0.39
Prevalence of learning disabilities, %	0 *	0 *	0	0	**Chi: 0.032**
Prevalence of liver disease, %	1	1 *	0 **	1 *	**Chi: <0.000**
Prevalence of mental health conditions, %	2 *	3 *	2 *	5 ***	**Chi: <0.000**
Prevalence of paraplegia, %	1 **	3 **	1 **	3 **	**Chi: <0.000**
Prevalence of peptic ulcer, %	4 *	4 *	2 **	3	**Chi: <0.000**
Prevalence of peripheral vascular disease, %	8 ***	11 ***	4 **	6 **	**Chi: <0.000**
Prevalence of renal disease, %	23 *	23 *	24 *	18 ***	**Chi: <0.000**
Prevalence of rheumatic disease, %	10 **	8 *	6 *	5 **	**Chi: <0.000**

****: Significantly different from all 3 other clusters; **: significantly different from 2 other clusters; *: significantly different from 1 other clusters; all at 0.05/4=0.0125 significance level (Bonferroni adjustment)*

The clusters demonstrate a great variation in future care utilisation within the high-risk stratum (see figure 1). Emergency care utilisation, which defines high-risk patients, is high for all clusters. Nevertheless, clusters 1 and 3 have emergency care utilisation rates that lie closer to the medium risk stratum than the high-risk average. Nonemergency hospitalisations and outpatient attendances for clusters 3 and 4 are at or even below the medium risk rate. GP care on the other hand is more homogenous, with the rates for each cluster close to the high-risk average.

While for each care setting there exist high and low utilisation clusters, they are not consistently the same clusters. Each cluster has a unique pattern of utilisation rates (see figure 2). Cluster 1 has high utilisation across most care types, with the exception of emergency care. Cluster 4 has the opposite pattern, with high emergency care use but low utilisation of other care types. Clusters 2 and 3 have high and low utilisations across all settings, respectively. The differences between the clusters are strongest for historic care utilisation, upon which the cluster analysis is based. However, each cluster exhibits the same pattern of utilisation in 2012.

DISCUSSION

Principle findings

The low, medium and high risk strata broadly correlate with care utilisation. For all care settings, the high-risk stratum has the highest historic and future utilisation. However, this study shows that, within the high-risk stratum, there is significant variation in care needs across the care continuum. The high-risk group can be split into four segments with different care utilisation rates, characteristics and care priorities.

Comparing historic and future utilisation for the four clusters, similar patterns can be observed, indicating that cluster analysis of historic data can help predict future needs. However, future utilisation rates were closer to the group mean for all clusters and all care settings than historic rates. This can be at least partially explained by regression to the mean (RTM), which is known to affect care utilisation predictions.[12 37 38] RTM describes the phenomenon where exceptionally high or low observations tend to be followed by less extreme observations in repeated measurements.[39] This effect is compounded if subjects are stratified based on baseline measurements, which is the case when patients are clustered based on their 2008-2011 utilisation.

Comparison to previous studies

This study shows that, while integrated care and case management initiatives often are indiscriminately aimed at high-risk patients, the actual needs of these patients vary widely. Many studies have discussed how best to identify,[13 14 40 41] or care for,[6 8 10 11 37 42] the high-risk population, but few have used data analysis to better understand different types of high-risk patients.

A major strength of this study is its reliance on data from both primary and acute care, to create a more comprehensive picture of care needs. While some risk prediction models, such as the Combined Predictive Model, include utilisation of non-acute care settings as predictor variables,[15] this detail is lost in the final risk score and the stratification. An utilisation-based segmentation analysis, as demonstrated in this study, can be used to bring out this detail.

Limitations and future research

While both primary and secondary care data were used in this study to understand care needs across the continuum, the picture is still incomplete. No patient-level linked data was available on utilisation of the A&E department, mental health, community and social care, and these were therefore left out of scope. This is an important limitation, as many initiatives will require integration of these settings. Future research should be done using more extensive datasets where these are available.

Another limitation is that the population used in this study is a random sample of patients in England. Local populations may see different sizes or types of segments within their risk strata. Moreover, this study uses a custom risk prediction algorithm. If providers are using a specific risk model, they are encouraged to replicate the analysis using their own population data and risk strata.

Implications for integrated care

Segmenting the high-risk stratum using cluster analysis can help tailor and target integrated care programmes. For example, cluster 1 uses relatively little emergency care, but has a high utilisation of nonemergency and outpatient care. Patients in this segment may not be the best target for primary care-led interventions aimed at reducing emergency hospitalisations, as their overall usage of emergency care is low and they may already be under management of a specialist.

Cluster 2 has the highest utilisation rates, the highest risk score and the most LTCs. Surprisingly; this segment is also the youngest of the four, with an average age of 67. Overall high care utilisation makes this cluster a worthwhile target for interventions aimed at reducing care use. As patients in this cluster have extensive care needs across different settings, they would likely benefit from care coordination and case management initiatives.

Cluster 3 is at 83 years the oldest segment. Despite their old age, disease prevalence among the patients in this cluster is generally lower. This is reflected in their lower than average care use across all settings. This segment shows that while interventions often focus on elderly patients,[6 37 43] this population group does not necessarily have the highest care usage.

Cluster 4 has one of the highest utilisation rates for emergency care, combined with a lower use of all other care services. Even GP care, which varies little for the other clusters, is below average for this group. This could indicate a lack of preventative primary care: patients in this cluster have on average 1.7 LTCs, but their low usage of primary care could be causing complications which require emergency care. This would make cluster 4 a prime target for enhances services and primary care-led interventions focused on preventing complications and emergency hospitalisations.

CONCLUSION

This paper shows that a high risk of emergency hospitalisation is not unequivocally linked to high overall care needs, or a particular pattern of care use across other care settings. While risk stratification based on emergency hospitalisation can predict general care utilisation rates, within the high-risk stratum there exist four very different patient types. Cluster analysis can enhance risk stratification by identifying groups of high-risk patients with unique care patterns across the care continuum, around which integrated care programmes can be designed.

STATEMENTS

Database: This study is based on data from the Clinical Practice Research Datalink obtained under license from the UK Medicines and Healthcare Products Regulatory Agency. However, the interpretation and conclusions contained in the study are those of the authors alone.

Data sharing: Technical appendix available in supplementary files, statistical code available from the corresponding author. No additional data available.

Declaration of competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: grants from Peter Sowerby Foundation, during the conduct of the study; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethics approval: No ethics approval was required

Funding: This study was partially funded by the Soweby eHealth Forum, sponsored by the Peter Sowerby Foundation. The funder had no role in the study design or analysis, or in the drafting and submission of this paper. The researchers worked independent from the funders.

Contributors: SV designed the study, created the database, analysed the data, and drafted and revised the paper. She is guarantor. EM contributed to the design of the

study, analysed the results and revised the draft paper. AD contributed to the design of the study and revised the draft paper. All have approved the final version for publication.

FIGURE LEGENDS

Figure 1: Mean future care utilisation for the risk strata - High (H), Medium (M) and Low (L) - and the four high-risk clusters - 1, 2, 3 and 4.

Figure 2: Patterns of utilisation for the four high-risk clusters – Emergency care hospitalisations (Emg), Nonemergency hospitalisations (NonE), Outpatient attendances (OP) and GP visits (GP) versus the high-risk population mean

REFERENCES

1. Zulman DM, Pal Chee C, Wagner TH, et al. Multimorbidity and healthcare utilisation among high-cost patients in the US Veterans Affairs Health Care System. BMJ Open 2015;**5**(4).
2. Department of Health. Supporting People with Long Term Conditions. An NHS and Social Care Model to support local innovation and integration. Leeds: Department of Health, 2005.
3. NHS England. Using case finding and risk stratification: A key service component for personalised care and support planning. Leeds: NHS England, 2015.
4. Goodwin N, Curry N. Methods for predicting risk of emergency hospitalisation: promoting self-care and integrated service responses in the home to the most vulnerable. Int J Integr Care 2008;**8**(5).

5. Dueñas-Espín I, Vela E, Pauws S, et al. Proposals for enhanced health risk assessment and stratification in an integrated care scenario. BMJ Open 2016;**6**(4).
6. Roland M, Lewis R, Steventon A, et al. Case management for at-risk elderly patients in the English integrated care pilots: observational study of staff and patient experience and secondary care utilisation. Int J Integr Care 2012;**12**(5).
7. Lewis G. Case study: Virtual wards at Croydon Primary Care Trust. London: The King's Fund, 2006.
8. Lewis G, Bardsley M, Vaithianathan R, et al. Do 'virtual wards' reduce rates of unplanned hospital admissions, and at what cost? A research protocol using propensity matched controls. Int J Integr Care 2011;**11**:e079.
9. NHS England. Enhanced service specification - Avoiding unplanned admissions: proactive case finding and patient review for vulnerable people. Leeds: NHS England, 2014.
10. Wallace E, Smith SM, Fahey T, et al. Reducing emergency admissions through community based interventions. Bmj 2016;**352**.
11. Lewis GH, Vaithianathan R, Wright L, et al. Integrating care for high-risk patients in England using the virtual ward model: lessons in the process of care integration from three case sites. Int J Integr Care 2013;**13**(4).
12. Lewis G. Next Steps for Risk Stratification in the NHS. London: NHS England, 2015.
13. Billings J, Blunt I, Steventon A, et al. Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30). BMJ Open 2012;**2**(4).
14. Billings J, Dixon J, Mijanovich T, et al. Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. Br Med J 2006;**333**:327.
15. Wennberg D, Siegel M, Darin B, et al. Combined Predictive Model - Final report & technical documentation. London: Health Dialog, King's Fund and New York University, 2006.
16. Thomson K, Lewis G. Information Governance and Risk Stratification: Advice and Options for CCGs and GPs. London: NHS England, 2013.
17. Lewis G, Curry N, Bardsley M. Choosing a Predictive Risk Model: A guide for commissioners in England. London: Nuffield Trust, 2011.
18. Berwick DM, Nolan TW, Whittington J. The Triple Aim: Care, Health, And Cost. Health Aff (Millwood) 2008;**27**(3):759-69.
19. The King's Fund. Predictive risk Project - Literature Review. London: The King's Fund, 2005.
20. Lewis G, Kirkham H, Duncan I, et al. How Health Systems Could Avert 'Triple Fail' Events That Are Harmful, Are Costly, And Result In Poor Patient Satisfaction. Health Aff (Millwood) 2013;**32**(4):669-76.

21. NHS Health and Social Care Information Centre. Numbers of Patients Registered at a GP Practice - July 2015. Leeds: NHS Health and Social Care Information Centre, 2015.

22. Bardsley M, Billings J, Dixon J, et al. Predicting who will use intensive social care: case finding tools based on linked health and social care data. Age Ageing 2011;**40**:265-70.

23. IBM SPSS Statistics for Macintosh, Version 23.0 [program]. Armonk, NY: IBM Corp, 2015.

24. Han J, Kamber M. *Data Mining: Concepts and Techniques*. 1st ed. San Diego, CA: Academic Press, 2001.

25. Stata Statistical Software: Release 14. [program]. College Station, TX: StataCorp LP, 2015.

26. Armstrong JJ, Zhu M, Hirdes JP, et al. K-Means Cluster Analysis of Rehabilitation Service Users in the Home Health Care System of Ontario: Examining the Heterogeneity of a Complex Geriatric Population. Arch Phys Med Rehabil 2012;**93**(12):2198-205.

27. Coste J, Bouyer J, Fernandez H, et al. A population-based analytical approach to assessing patterns, determinants, and outcomes of health care with application to ectopic pregnancy. Med Care 2000;**38**(7):739-49.

28. Cryer PC, Saunders J, Jenkins LM, et al. Clusters within a general adult population of alcohol abstainers. International Journal of Epidemiology 2001;**30**(4):756-65.

29. Kendig H, Mealing N, Carr R, et al. Assessing patterns of home and community care service use and client profiles in Australia: a cluster analysis approach using linked data. Health & Social Care in the Community 2012;**20**(4):375-87.

30. Pud D, Ben Ami S, Cooper BA, et al. The Symptom Experience of Oncology Outpatients Has a Different Impact on Quality-of-Life Outcomes. J Pain Symptom Manage 2008;**35**(2):162-70.

31. Everitt BS, Landau S, Leese M, et al. *Cluster analysis*. 5th ed. Chichester: John Wiley & Sons, 2011.

32. StataCorp. Stata 14 Cluster Stop reference manual. College Station, TX: Stata Press, 2015.

33. Ng SK, Holden L, Sun J. Identifying comorbidity patterns of health conditions via cluster analysis of pairwise concordance statistics. Stat Med 2012;**31**(27):3393-405.

34. Chan M, Zhu MX. Investigating the health profile of Macau Chinese. J Clin Nurs 2008;**17**(11C):352-61.

35. Borglin G, Jakobsson U, Edberg Ak, et al. Older people in Sweden with various degrees of present quality of life: their health, social support, everyday activities and sense of coherence. Health & Social Care in the Community 2006;**14**(2):136-46.

36. Kansagara D, Englander H, Salanitro A, et al. Risk prediction models for hospital readmission: A systematic review. JAMA 2011;**306**(15):1688-98.

37. Roland M, Abel G. Reducing emergency admissions: are we on the right track? Br Med J 2012;**345**(e6017).

38. Georghiou T, Blunt I, Steventon A, et al. Predictive risk and health care: An overview. London: Nuffield Trust, 2011.

39. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. International Journal of Epidemiology 2005;**34**(1):215-20.

40. Hao S, Jin B, Shin AY, et al. Risk prediction of emergency department revisit 30 days post discharge: a prospective study. PloS one 2014;**9**(11):e112944.

41. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. Bmj 2015;**351**.

42. Tanio C, Chen C. Innovations At Miami Practice Show Promise For Treating High-Risk Medicare Patients. Health Aff (Millwood) 2013;**32**(6):1078-82.

43. Alderwick H, Ham C, Buck D. Population health systems: Going beyond integrated care. London: The King's Fund, 2015.

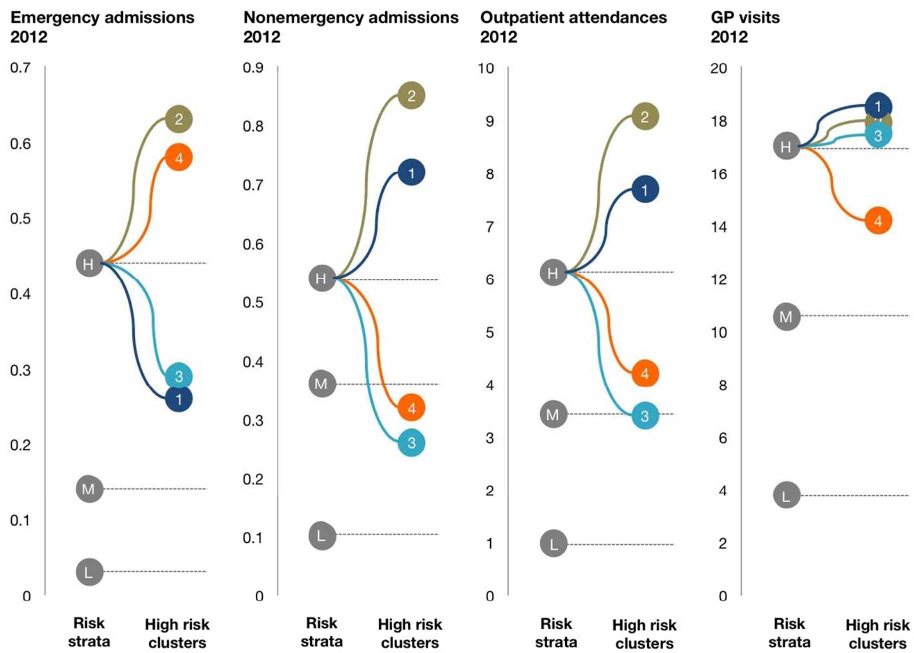
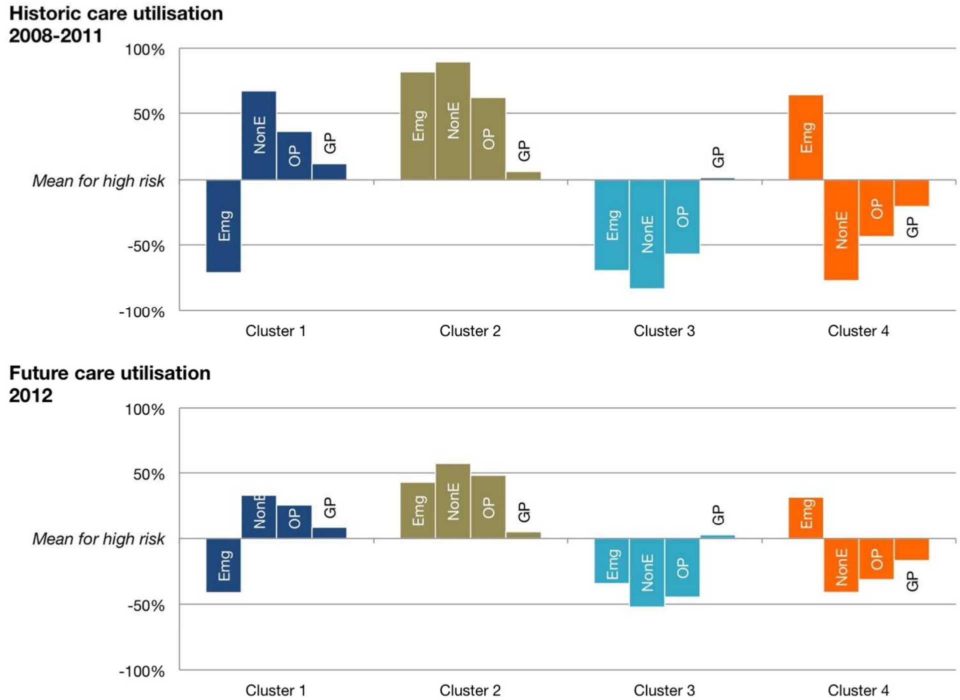


Figure 1: Mean future care utilisation for the risk strata - High (H), Medium (M) and Low (L) - and the four high-risk clusters - 1, 2, 3 and 4.
figure 1
508x381mm (72 x 72 DPI)



Patterns of utilisation for the four high-risk clusters – Emergency care hospitalisations (Emg), Nonemergency hospitalisations (NonE), Outpatient attendances (OP) and GP visits (GP) versus the high-risk population mean

figure 2

508x381mm (72 x 72 DPI)

Appendix 1

Variables included in various risk scores and variables selected for our model

Variables considered in hospital admission risk studies	Number of studies out of 30 including variable in final model[1]	Variable in PARR[2]	Variable in PARR-30[3]	Variable in Combined Predictive Model[4] (*selected variables*)	Included in initial model / *included in final model after backwards elimination*
Morbidities	Medical diagnoses or comorbidity indices: 24	Cerebrovascular disease			Any diagnosis of cerebrovascular disease in 2008-2011 (in primary or secondary care)
		Chronic obstructive pulmonary disease	Chronic pulmonary disease	COPD	***Any diagnosis of COPD in 2008-2011 (in primary or secondary care)***
				Asthma (only considered in LTC counts)	Any diagnosis of Asthma in 2008-2011 (in primary or secondary care)
		Connective tissue disease/rheumatoid arthritis			***Any diagnosis of Rheumatic disease in 2008-2011 (in primary or secondary care)***
		Developmental disability			Any diagnosis of Learning disability in 2008-2011 (in primary or secondary care)
		Diabetes	Diabetes with chronic complications	*Diabetes (only considered in LTC counts)*	Any diagnosis of Diabetes in 2008-2011 (in primary or

				secondary care)
	Ischaemic heart disease		*CAD (only considered in LTC counts)*	***Any diagnosis of Ischaemic heart disease in 2008-2011 (in primary or secondary care)***
	Peripheral vascular disease	Peripheral vascular disease		Any diagnosis of Peripheral vascular disease in 2008-2011 (in primary or secondary care)
	Renal failure	Renal disease		***Any diagnosis of Renal disease in 2008-2011 (in primary or secondary care)***
	Sickle cell disease			
		Metastatic cancer with solid tumour	*Cancer (only considered in LTC counts)*	***Any diagnosis of Cancer in 2008-2011 (in primary or secondary care)***
		Other malignant cancer		
		Congestive heart failure	*CHF (only considered in LTC counts)*	Any diagnosis of Congestive heart failure in 2008-2011 (in primary or secondary care)
		Moderate/severe liver disease		Any diagnosis of Liver disease in 2008-2011 (in primary or secondary care)
		Other liver disease		
		Haemiplegia or paraplegia		Any diagnosis of Paraplegia in 2008-2011 (in primary or secondary care)

			Dementia		Any diagnosis of Dementia in 2008-2011 (in primary or secondary care)
				Hypertension (only considered in LTC counts)	
		Diagnostic cost groups/hierarchical condition category			
				1 LTC	***Flag if the sum of conditions listed is 0, 1 or 2 or more***
				2+ LTCS	
Mental health morbidities	Alcohol or substance use: 11	Alcohol related diagnosis		Psychoactive substance abuse	
	Mental illness: 9			Psychotic disorder	
				Inpatient admission with diagnosis of mental illness	***Any diagnosis of Mental health disorder in 2008-2011 (in primary or secondary care)***
				Depression (only as included in LTC counts)	
Prior use of medical services	Hospitalisations: 14	Previous admission for respiratory infection			
		Previous admission for a reference condition			
		Number of emergency admissions in previous 90, 180 and 365 days	Whether there had been a prior emergency hospital discharge in the past 30	[Combinations of] 1, 1+, 2, 2+, 3+ emergency admissions in last 30, 30 to 90, 90 to 180, 180 to	***Number of emergency admissions in 2011*** &

			days	365, 365 to 730 days	***Number of emergency admissions over 2008-2011***
		Average number of episodes per spell for emergency admissions		Average number of episodes per spell for emergency admissions >=3	
			Whether the current admission was an emergency admission		
		Total number of previous emergency admissions in previous three years	Number of emergency hospital discharges in the last year		
		Number of non-emergency admissions in previous 365 days			***Number of non-emergency admissions over 2008-2011***
	Emergency department visits: 4			A&E visits and investigations	
	Clinic visits or missed visits: 3	Number of different treatment specialists seen			
				[Combinations of] 1, 1-5, 2, 3+, 6-10, 11+ out-patient specialty visits in last 30, 30 to 90, 365 to 730 days	***Number of outpatient visits over 2008-2011***
	Index hospital length of stay: 4				
	Other			Polypharmacy: 1-4 unique drugs in any month (last 0 to 90 days);	***Number of GP visits over 2008-2011 (including home***

				5-9; 10+	***visits)***
Sociodemographic factors	Age: 19	Age 65-74 or age 75+	Age squared	Age band (0-4, 15-39, 40-59, 5 year age bands, 85+)	***5-year age bands*** & Over 75 flag
	Sex: 15	Sex		Gender	***Gender***
	Race/ ethnicity: 7	Ethnicity			
Social determinants of health	SES, income and employment: 5		Index of multiple deprivation band for the place of residence		***Townsend score (5 groups)***
	Insurance status: 6				
	Education: 0				
	Marital status and people in household: 4				
	Social support: 2				
	Access to care: 5				
	Discharge location: 2				
Hospital specific metrics	*Not included in review*	Observed:expected ratio for practice style sensitive admissions in ward of residence			
		Observed:expected ratio for rate of readmissions for hospitals of current admission	Hospital-specific variable		
Illness severity	Severity index: 1				
	Laboratory findings: 4				

	Other: 4	
Overall health and function	Functional status, ADL: 2	
	Self-rated health, QOL: 3	
	Cognitive impairment: 7	
	Visual/hearing impairment: 1	

1. Kansagara D, Englander H, Salanitro A, et al. Risk prediction models for hospital readmission: A systematic review. JAMA 2011;**306**(15):1688-98.
2. Billings J, Dixon J, Mijanovich T, et al. Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. Br Med J 2006;**333**:327.
3. Billings J, Blunt I, Steventon A, et al. Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30). BMJ Open 2012;**2**(4).
4. Wennberg D, Siegel M, Darin B, et al. Combined Predictive Model - Final report & technical documentation. London: Health Dialog, King's Fund and New York University, 2006.

BMJ Open

Enhancing risk stratification for use in integrated care - A cluster analysis of high-risk patients in a retrospective cohort study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2016-012903.R1
Article Type:	Research
Date Submitted by the Author:	23-Sep-2016

Complete List of Authors:	Vuik, Sabine; Imperial College London, Institute of Global Health Innovation Mayer, Erik; Imperial College London, Dept. of Biosurgery and Surgical Technology Darzi, Ara; Imperial College London, Institute of Global Health Innovation

Primary Subject Heading:	Health services research
Secondary Subject Heading:	Research methods, Patient-centred medicine, Evidence based practice, General practice / Family practice
Keywords:	Risk management < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Organisation of health services < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, STATISTICS & RESEARCH METHODS

Enhancing risk stratification for use in integrated care - A cluster analysis of high-risk patients in a retrospective cohort study

Sabine I Vuik, Erik Mayer, Ara Darzi

Sabine I Vuik, Policy Fellow, Institute of Global Health Innovation

Imperial College, 10th Floor, St Mary's Hospital, Praed Street, London, W1 2NY, UK

Erik Mayer, Clinical Senior Lecturer, Department of Surgery

Imperial College, 10th Floor, St Mary's Hospital, Praed Street, London, W1 2NY, UK

Ara Darzi, Professor in Surgery, Department of Surgery

Imperial College, 10th Floor, St Mary's Hospital, Praed Street, London, W1 2NY, UK

Correspondence to: Sabine Vuik s.vuik@imperial.ac.uk, +44(0) 795 714 0479

Keywords: risk prediction, integrated care, care utilisation, emergency hospitalisation, high-risk patients

Word count (excl. title page, abstract, references, tables and figures): 2,456

ABSTRACT

Objective: To show how segmentation can enhance risk stratification tools for integrated care, by providing insight into different care utilisation patterns within the high-risk population.

Design: A retrospective cohort study. A risk score was calculated for each person using a logistic regression, which was then used to select the top 5% high-risk individuals. This population was segmented based on utilisation of different care settings using a k-means cluster analysis. Data from 2008 to 2011 was used to create the risk score and segments, while 2012 data was used to understand the predictive abilities of the models.

Setting and participants: Data was collected from administrative datasets covering primary and secondary care for a random sample of 300,000 English patients.

Main measures: The high-risk population was segmented based on their utilisation of four different care settings: emergency acute care, elective acute care, outpatient care and GP care.

Results: While the risk strata predicted care utilisation at a high level, within the high-risk population utilisation varied significantly. Four different groups of high-risk patients could be identified. These four segments had distinct utilisation patterns across care settings, reflecting different levels and types of care needs. The 2008-2011 utilisation patterns of the four segments were consistent with the 2012 patterns.

Discussion: Cluster analyses revealed that the high-risk population is not homogeneous, as there exist four groups of patients with different needs across the care continuum. Since the patterns were predictive of future care use, they can be used to develop integrated care programmes tailored to these different groups.

Conclusions: Utilisation-based segmentation augments risk stratification by identifying patient groups with different care needs, around which integrated care programmes can be designed.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- This study uses a large dataset containing patient-level linked primary and secondary care administrative data
- Rather than focusing only on emergency care, this study looks at patterns of utilisation across different care settings to support the development of integrated care programmes

- Where previous studies have focused on how to identify or manage high-risk patients, this study explores the different patient groups within the high-risk stratum
- The data used was for a random sample of English patients, and may not reflect local trends
- No data was available in linked format for other care settings, such as A&E, mental health, community and social care

BACKGROUND

In healthcare, a small number of patients accounts for a disproportionally large share of utilisation.[1 2] Identifying and targeting this group can be done through risk stratification. Risk stratification divides a population based on different levels of risk of a specific outcome, and is often presented as a core process to achieve integrated, personalised care.[3-5] For each stratum, a tailored care model can be developed which addresses the specific needs of the patients. Many of the interventions for high-risk patients are primary care-led integrated care programmes, like virtual wards, case management, and enhanced services and access. [4 6-11]

Risk stratification methods often focus on predicting emergency hospitalisations.[3 12-15] Unplanned hospitalisations, including readmissions, are chosen because they are costly for a health system, may indicate low quality care, and have a negative impact on patient experience.[16 17] As such, unplanned hospitalisations are reflective of all elements of the triple aim of healthcare – quality of care, patient experience and cost[18] – and can be considered a 'triple fail event'.[16] Moreover, since preventing emergency hospitalisations to the acute setting requires effective primary care, they are also an important metric for integrated care.[19]

However, risk stratification based on emergency hospitalisations has important limitations. Firstly, this approach only looks at one element of care. While the risk of an emergency hospitalisation can be expected to correlate with overall use of emergency acute care, utilisation of other care services may vary. A patient with an emergency hospitalisation may be under treatment with a specialist; or regularly visit a general practitioner (GP); or not access ambulatory care at all. In order to design effective integrated care programmes that link up the appropriate care providers, understanding care use across all settings is crucial.

Secondly, detailed information on the characteristics of the high-risk patients, such as age, morbidities and socio-economic status, is lost in the final risk score. All patients who end up in the top stratum have high risk scores, but the factors driving this high score can be very different. When developing interventions, these should be taken into account to understand which patients are most likely to respond to different interventions.[12 20]

The aim of this study is to show how utilisation-based segmentation can enhance risk stratification tools used for integrated care by, firstly, taking into account care utilisation across multiple care settings and, secondly, providing insight into the characteristics of different patient groups within the high-risk stratum.

METHODS

Study design

To show how segmentation can augment risk stratification, we applied both methods to a large patient database. We first trained a risk prediction model to generate risk scores for each patient. Based on these risk scores, we identified the high-risk patient population. In this group we applied a cluster analysis to a range of different utilisation variables. The different clusters were analysed and profiled to understand the different patient types that exist within a high-risk group.

The analyses were conducted for hypothetical "historic" (2008-2011) and "future" (2012) datasets. The historic dataset reflects the information that would be available to healthcare professionals conducting risk stratification and cluster analysis at the end of 2011, while the future dataset was used to understand how accurately the models predicted actual utilisation in the following year.

Software

STATA (version 14)[21] was used to perform the cluster analyses and calculate the pseudo-F statistics. For all other analyses, including the risk prediction, SPSS (version 23)[22] was used.

Data

A dataset covering primary and secondary care use for a random sample of 300,000 English patients was constructed from Clinical Practice Research Datalink (CPRD) and Hospital Episode Statistics (HES) data (CPRD ISAC approval under protocol 14_211R). Patients were eligible for inclusion if they were registered with a CPRD-participating GP practice during the entire study period of 2008 up to and including 2012, and if their HES records could be linked to CPRD. Other than those two criteria, the sample was entirely random. The CPRD dataset is broadly representative of the age, sex and ethnicity composition of the UK population.[23] In England, Clinical Commissioning Groups (CCG) are responsible for the planning and commissioning of care for local populations. The sample size in this study was set at 300,000, which is similar to the population of a CCG in the 75th percentile,[24] to reflect a typical local population in England.

The final dataset included patient demographics, long-term condition (LTC) diagnoses and utilisation variables. We selected four high-level utilisation variables for the cluster analysis of high-risk patients: inpatient emergency hospitalisations, inpatient nonemergency hospitalisations, outpatient attendances and GP visits. These utilisation variables were used to reflect different care settings that may be incorporated in integrated care models. For the cluster analysis, the utilisation variables were log-normalised and standardised to reduce the impact of outliers and give equal weight to each variable.

Risk stratification

We calculated our own risk prediction score, reflecting predictor variables used in Patients at Risk of Re-hospitalisation (PARR) tool, the Combined Predictive Model and other commonly used risk prediction algorithms. The risk model was trained to predict emergency hospitalisations in 2012, using a stepwise logistic regression.[14 25] The number of emergency hospitalisations in 2011 was included as one of the predictor variables, as well as a range of other variables used in previous risk models,[13-15 26] as detailed in appendix 1. The logistic regression on the training set excluded a number of diagnosis variables after step-wise elimination, as well as the 75+ flag.

To validate the model, a split sample validation method was used. Using the random sample function of SPSS, half of the sample was defined as the training set and the other half as the test set. Applying the risk model to the test set, the area under the Receiver Operator Curve (ROC) was 0.75. This is in line with other models predicting emergency hospitalisations, which range from 0.55 to 0.83.[13 26] The test population was stratified into three groups, which are comprised of the top 5% highest risk patient ("High risk"), the top 5-20% ("Medium risk") and the remaining 80% of the population ("Low risk"), in accordance with general risk stratification practice.[2 15 17]

Segmentation

For the segmentation analysis the k-means algorithm was used to cluster the patients based on their historic utilisation. This method was selected as it is efficient and produces roughly similar sized segments.[27] Clustering solutions ranging from 2 to 8 clusters were explored for the high-risk stratum. To identify the optimal number of clusters, the Pseudo-F statistic was calculated for all the clustering solutions using STATA. This statistic is commonly used in healthcare clustering studies,[28-32] and is one of the best criteria to determine the number of clusters.[33] It compares the between-cluster to the within-cluster sum-of-squares, and a large Pseudo-F statistic indicates distinct clusters.[34] In addition, the different clustering solutions were also explored using Ward's linkage clustering and post-hoc analysis, as detailed in appendix 2. Both the k-means and Ward's clustering analyses used the Euclidian distance measure.

The clusters were evaluated based on their validity, through statistical test confirming the differences between clusters, and their stability, by comparing future care utilisation of each cluster to the historic pattern.

Analysis

To create profiles for the segments, the utilisation variables as well as demographic characteristics were analysed to see if they differed significantly across segments. For the non-Normal utilisation and LTCs count variables, a Kruskal-Wallis test was used. For the continuous age and risk score variables an ANOVA test was used, and for the binary morbidity variables and the 2012 emergency hospitalisation flag a Chi square test. Where these tests found significant variation across segments, the results were then explored pair-wise between segments to identify which segment or segments were significantly different from others. For this, Mann-Whitney U tests, Student t-tests, and z-tests were used, respectively. To account for the multiplicity problem that occurs when performing multiple tests, the Bonferroni method was used to adjust the significance level.[35-37]

RESULTS

The final dataset contained 298,111 people with a complete record across the variables, of which 149,320 observations were allocated to the test set used for the analyses below. When the population was stratified based on risk, predictive variables such as age, long-term conditions and historic care utilisation were all found to increase with each risk stratum (see table 1). In addition to historic utilisation, future utilisation of all care types also increased for the high-risk stratum.

Table 1: Strata characteristics

	High risk	*Medium risk*	*Low risk*	***Total population***

Number of people	7,466	22,398	119,456	**149,320**
Predicted proportion with any emergency hospitalisations in 2012 (based on the average risk score)	27%	9%	3%	**5%**
Actual proportion with any emergency hospitalisations in 2012	27%	11%	3%	**5%**
Age at end of study period, mean	75	65	40	**45**
Number of long-term conditions, median (Interquartile Range/IQR)	2 (1 to 2)	1 (0 to 1)	0 (0 to 0)	**0 (0 to 0)**
Number of emergency hospitalisations over 2008-2011, median (IQR)	1 (1 to 3)	0 (0 to 1)	0 (0 to 0)	**0 (0 to 0)**
Number of nonemergency hospitalisations over 2008-2011, median (IQR)	1 (0 to 3)	1 (0 to 2)	0 (0 to 0)	**0 (0 to 1)**
Number of outpatient attendances over 2008-2011, median (IQR)	16 (8 to 30)	8 (2 to 16)	1 (0 to 4)	**1 (0 to 6)**
Number of GP visits over 2008-2011 median (IQR)	55 (35 to 82)	34 (22 to 51)	10 (4 to 20)	**13 (6 to 27)**
Number of emergency hospitalisations in 2012, median (IQR)	0 (0 to 1)	0 (0 to 0)	0 (0 to 0)	**0 (0 to 0)**
Number of nonemergency hospitalisations in 2012, median (IQR)	0 (0 to 1)	0 (0 to 0)	0 (0 to 0)	**0 (0 to 0)**
Number of outpatient attendances in 2012, median (IQR)	4 (1 to 8)	1 (0 to 4)	0 (0 to 1)	**0 (0 to 2)**
Number of GP visits in 2012, median (IQR)	13 (7 to 22)	8 (5 to 14)	2 (0 to 5)	**3 (1 to 7)**

For the high-risk population, k-means cluster analyses were performed for 2- to 8-clusters and the pseudo-F statistics was obtained for each solution. A peak was observed around the 3- and 4-cluster solutions. Exploring these two sets of clusters, the 4-cluster solution included an additional, contrasting utilisation pattern and was therefore selected.

The cluster analysis aims to optimise the distance between groups for the clustering variables, and statistical tests confirm that historic utilisation is significantly different across segments (see table 2). In addition, non-clustering variables, including future utilisation, age, number of long-term conditions and most disease prevalence variables, also differ significantly across the clusters.

Table 2: Clusters within the high-risk population

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	ANOVA/ Kruskal-Wallis/ Chi square test
Clustering variables					
Number of emergency hospitalisations over 2008-2011, median (IQR)	1 (0 to 1) **	3 (2 to 4) ***	1 (0 to 1) **	3 (2 to 4) ***	**KW: <0.000**
Number of nonemergency hospitalisations over 2008-2011, median (IQR)	3 (2 to 5) ***	3 (2 to 5) ***	0 (0 to 1) ***	0 (0 to 1) ***	**KW: <0.000**
Number of outpatient attendances over 2008-2011, median (IQR)	24 (16 to 38) ***	29 (18 to 46) ***	7 (3 to 13) ***	10 (5 to 18) ***	**KW: <0.000**
Number of GP visits over 2008-2011, median (IQR)	61 (43 to 90) ***	57 (40 to 86) ***	55 (35 to 82) ***	42 (26 to 65) ***	**KW: <0.000**
Post-hoc analysis of other variables					
Number of people	1967	1807	1831	1861	
Predicted proportion with any emergency hospitalisations in 2012 (based on average risk score), %	21 ***	38 ***	20 ***	31 ***	**AN: <0.000**
Actual proportion with any emergency hospitalisations in 2012, %	19 **	35 **	21 **	34 **	**Chi: <0.000**
Age at end of study period, mean	79 ***	67 ***	83 ***	71 ***	**AN: <0.000**
Number of long-term conditions, median (IQR)	2 (1 to 3) **	2 (1 to 3) **	1 (1 to 2) ***	1 (1 to 2) ***	**KW: <0.000**
Number of emergency hospitalisations in 2012, median (IQR)	2 (1 to 3) **	2 (1 to 3) **	1 (1 to 2) **	1 (1 to 2) **	**KW: <0.000**
Number of nonemergency hospitalisations in 2012, median (IQR)	0 (0 to 0) **	0 (0 to 1) **	0 (0 to 0) ***	0 (0 to 1) ***	**KW: <0.000**
Number of outpatient attendances in 2012, median (IQR)	0 (0 to 1) ***	0 (0 to 1) ***	0 (0 to 0) ***	0 (0 to 0) ***	**KW: <0.000**
Number of GP visits in 2012,	5 (2 to 10) ***	6 (3 to 11) **	2 (0 to **	2 (0 to 5) ***	**KW: <0.000**

median (IQR)			4)		
Prevalence of acute myocardial infarction, %	15 ***	23 ***	10 ***	19 ***	**Chi: <0.000**
Prevalence of asthma, %	28 *	26	24 *	25	**Chi: 0.028**
Prevalence of cancer, %	26 ***	22 ***	8 ***	5 ***	**Chi: <0.000**
Prevalence of cerebrovascular disease, %	9 **	15 **	10 **	18 **	**Chi: <0.000**
Prevalence of congestive heart failure, %	8 ***	13 **	5 ***	13 **	**Chi: <0.000**
Prevalence of COPD, %	18 *	17 *	13 ***	18 *	**Chi: <0.000**
Prevalence of dementia, %	3 **	3 **	5 **	7 **	**Chi: <0.000**
Prevalence of diabetes, %	28 **	22 **	28 **	22 **	**Chi: <0.000**
Prevalence of HIV/AIDS, %	0	0	0	0	Chi: 0.39
Prevalence of learning disabilities, %	0 *	0 *	0	0	**Chi: 0.032**
Prevalence of liver disease, %	1	1 *	0 **	1 *	**Chi: <0.000**
Prevalence of mental health conditions, %	2 *	3 *	2 *	5 ***	**Chi: <0.000**
Prevalence of paraplegia, %	1 **	3 **	1 **	3 **	**Chi: <0.000**
Prevalence of peptic ulcer, %	4 *	4 *	2 **	3	**Chi: <0.000**
Prevalence of peripheral vascular disease, %	8 ***	11 ***	4 **	6 **	**Chi: <0.000**
Prevalence of renal disease, %	23 *	23 *	24 *	18 ***	**Chi: <0.000**
Prevalence of rheumatic disease, %	10 **	8 *	6 *	5 **	**Chi: <0.000**

****: Significantly different from all 3 other clusters; **: significantly different from 2 other clusters; *: significantly different from 1 other clusters; all at 0.05/4=0.0125 significance level (Bonferroni adjustment)*

The clusters demonstrate a great variation in future care utilisation within the high-risk stratum (see figure 1). Emergency care utilisation, which defines high-risk patients, is high for all clusters. Nevertheless, clusters 1 and 3 have emergency care utilisation rates that lie closer to the medium risk stratum than the high-risk average. Nonemergency hospitalisations and outpatient attendances for clusters 3 and 4 are at or even below the medium risk rate. GP care on the other hand is more homogenous, with the rates for each cluster close to the high-risk average.

While for each care setting there exist high and low utilisation clusters, they are not consistently the same clusters. Each cluster has a unique pattern of utilisation rates (see figure 2). Cluster 1 has high utilisation across most care types, with the exception of emergency care. Cluster 4 has the opposite pattern, with high emergency care use but low utilisation of other care types. Clusters 2 and 3 have high and low utilisations across all settings, respectively. The differences between the clusters are strongest for historic care utilisation, upon which the cluster analysis is based. However, each cluster exhibits the same pattern of utilisation in 2012.

DISCUSSION

Principle findings

The low, medium and high risk strata broadly correlate with care utilisation. For all care settings, the high-risk stratum has the highest historic and future utilisation. However, this study shows that, within the high-risk stratum, there is significant variation in care needs across the care continuum. The high-risk group can be split into four segments with different care utilisation rates, characteristics and care priorities.

Comparing historic and future utilisation for the four clusters, similar patterns can be observed, indicating that cluster analysis of historic data can help predict future needs. However, future utilisation rates were closer to the group mean for all clusters and all care settings than historic rates. This can be at least partially explained by regression to the mean (RTM), which is known to affect care utilisation predictions.[12 38 39] RTM describes the phenomenon where exceptionally high or low observations tend to be followed by less extreme observations in repeated measurements.[40] This effect is compounded if subjects are stratified based on baseline measurements, which is the case when patients are clustered based on their 2008-2011 utilisation.

Comparison to previous studies

This study shows that, while integrated care and case management initiatives often are indiscriminately aimed at high-risk patients, the actual needs of these patients vary widely. Many studies have discussed how best to identify,[13 14 41 42] or care for,[6 8 10 11 38 43] the high-risk population, but few have used data analysis to better understand different types of high-risk patients.

A major strength of this study is its reliance on data from both primary and acute care, to create a more comprehensive picture of care needs. While some risk prediction models, such as the Combined Predictive Model, include utilisation of non-acute care settings as predictor variables,[15] this detail is lost in the final risk score and the stratification. An utilisation-based segmentation analysis, as demonstrated in this study, can be used to bring out this detail.

Limitations and future research

While both primary and secondary care data were used in this study to understand care needs across the continuum, the picture is still incomplete. No patient-level linked data was available on utilisation of the Accident and Emergency (A&E) department, mental health, community and social care, and these were therefore left out of scope. This is an important limitation, as many initiatives will require integration of these settings. Future research should be done using more extensive datasets where these are available.

Another limitation is that the population used in this study is a random sample of patients in England. In this specific sample, the long-term condition prevalence was relatively low. This could be attributable to the fact that conditions were identified based on coded diagnoses in the administrative data rather than from disease registries, but it could also be a characteristic of our sample. Local populations may see different sizes or types of segments within their risk strata. Moreover, this study uses a custom risk prediction algorithm. If providers are using a specific risk model, they are encouraged to replicate the analysis using their own population data and risk strata.

Implications for integrated care

Segmenting the high-risk stratum using cluster analysis can help tailor and target integrated care programmes. For example, cluster 1 uses relatively little emergency care, but has a high utilisation of nonemergency and outpatient care. Patients in this segment may not be the best target for primary care-led interventions aimed at reducing emergency hospitalisations, as their overall usage of emergency care is low and they may already be under management of a specialist.

Cluster 2 has the highest utilisation rates, the highest risk score and the most LTCs. Surprisingly; this segment is also the youngest of the four, with an average age of 67. Overall high care utilisation makes this cluster a worthwhile target for interventions aimed at reducing care use. As patients in this cluster have extensive care needs across different settings, they would likely benefit from care coordination and case management initiatives.

Cluster 3 is at 83 years the oldest segment. Despite their old age, disease prevalence among the patients in this cluster is generally lower. This is reflected in their lower than average care use across all settings. This segment shows that while interventions often focus on elderly patients,[6 38 44] this population group does not necessarily have the highest care usage.

Cluster 4 has one of the highest utilisation rates for emergency care, combined with a lower use of all other care services. Even GP care, which varies little for the other clusters, is below average for this group. This could indicate a lack of preventative primary care: patients in this cluster have on average 1.7 LTCs, but their low usage of primary care could be causing complications which require emergency care. This would make cluster 4 a prime target for enhances services and primary care-led interventions focused on preventing complications and emergency hospitalisations.

However, it is important to note that the above implications are theoretical and have not been confirmed in practice. Future research is needed to translate the theoretical concepts presented in this paper into actionable information, including effective interventions and implementation.

CONCLUSION

This paper shows that a high risk of emergency hospitalisation is not unequivocally linked to high overall care needs, or a particular pattern of care use across other care settings. While risk stratification based on emergency hospitalisation can predict general care utilisation rates, within the high-risk stratum there exist four very different patient types. Cluster analysis can enhance risk stratification by identifying groups of high-risk patients with unique care patterns across the care continuum, around which integrated care programmes can be designed.

STATEMENTS

Database: This study is based on data from the Clinical Practice Research Datalink obtained under license from the UK Medicines and Healthcare Products Regulatory Agency. However, the interpretation and conclusions contained in the study are those of the authors alone.

Data sharing: Technical appendix available in supplementary files, statistical code available from the corresponding author. No additional data available.

Declaration of competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: grants from Peter Sowerby Foundation, during the conduct of the study; no financial relationships with any organisations that might have an interest in the submitted work in the previous three

years; no other relationships or activities that could appear to have influenced the submitted work.

Ethics approval: No ethics approval was required

Funding: This study was partially funded by the Sowerby eHealth Forum, sponsored by the Peter Sowerby Foundation. The funder had no role in the study design or analysis, or in the drafting and submission of this paper. The researchers worked independent from the funders.

Contributors: SV designed the study, created the database, analysed the data, and drafted and revised the paper. She is guarantor. EM contributed to the design of the study, analysed the results and revised the draft paper. AD contributed to the design of the study and revised the draft paper. All have approved the final version for publication.

FIGURE LEGENDS

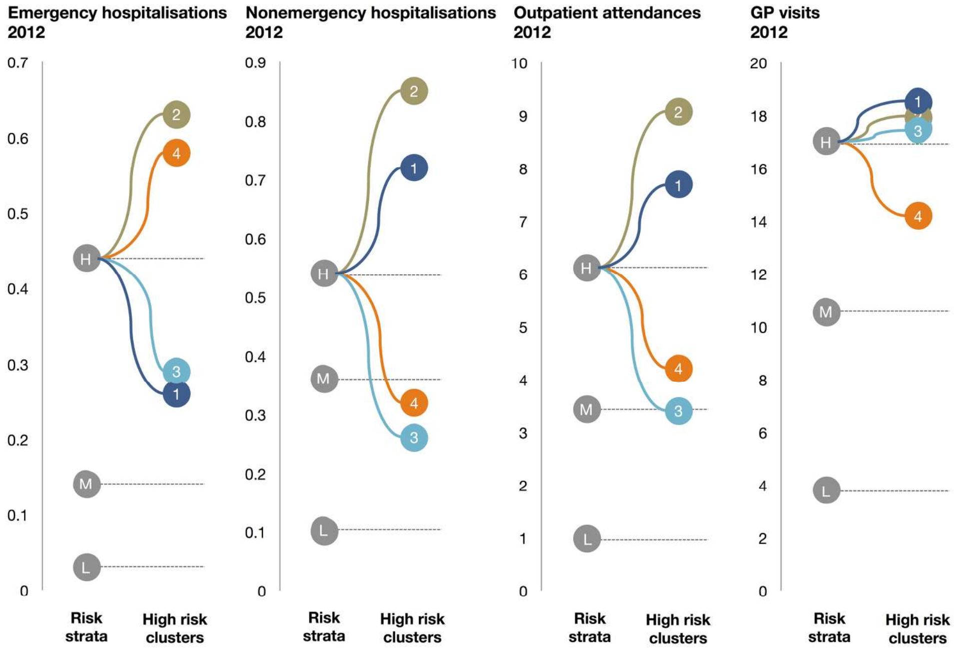
Figure 1: Mean future care utilisation for the risk strata - High (H), Medium (M) and Low (L) - and the four high-risk clusters - 1, 2, 3 and 4.

Figure 2: Patterns of utilisation for the four high-risk clusters – Emergency care hospitalisations (Emg), Nonemergency hospitalisations (NonE), Outpatient attendances (OP) and GP visits (GP) versus the high-risk population mean

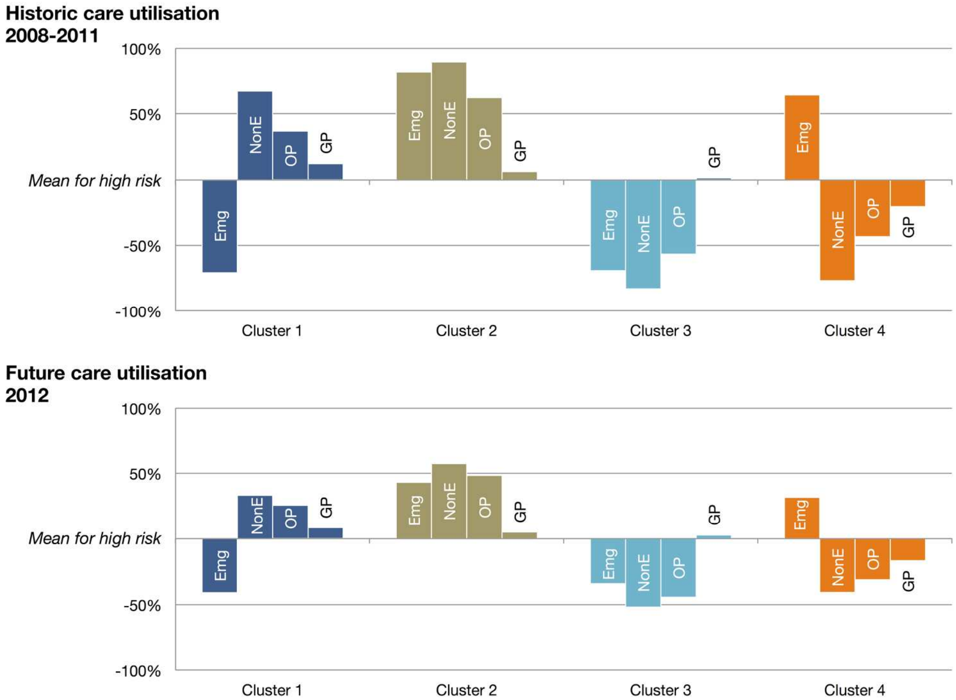
REFERENCES

1. Zulman DM, Pal Chee C, Wagner TH, et al. Multimorbidity and healthcare utilisation among high-cost patients in the US Veterans Affairs Health Care System. BMJ Open 2015;**5**(4).
2. Department of Health. Supporting People with Long Term Conditions. An NHS and Social Care Model to support local innovation and integration. Leeds: Department of Health, 2005.
3. NHS England. Using case finding and risk stratification: A key service component for personalised care and support planning. Leeds: NHS England, 2015.
4. Goodwin N, Curry N. Methods for predicting risk of emergency hospitalisation: promoting self-care and integrated service responses in the home to the most vulnerable. Int J Integr Care 2008;**8**(5).
5. Dueñas-Espín I, Vela E, Pauws S, et al. Proposals for enhanced health risk assessment and stratification in an integrated care scenario. BMJ Open 2016;**6**(4).
6. Roland M, Lewis R, Steventon A, et al. Case management for at-risk elderly patients in the English integrated care pilots: observational study of staff and patient experience and secondary care utilisation. Int J Integr Care 2012;**12**(5).
7. Lewis G. Case study: Virtual wards at Croydon Primary Care Trust. London: The King's Fund, 2006.
8. Lewis G, Bardsley M, Vaithianathan R, et al. Do 'virtual wards' reduce rates of unplanned hospital admissions, and at what cost? A research protocol using propensity matched controls. Int J Integr Care 2011;**11**:e079.
9. NHS England. Enhanced service specification: Avoiding unplanned admissions: proactive case finding and patient review for vulnerable people. Leeds: NHS England, 2014.
10. Wallace E, Smith SM, Fahey T, et al. Reducing emergency admissions through community based interventions. Br Med J 2016;**352**.
11. Lewis GH, Vaithianathan R, Wright L, et al. Integrating care for high-risk patients in England using the virtual ward model: lessons in the process of care integration from three case sites. Int J Integr Care 2013;**13**(4).
12. Lewis G. Next Steps for Risk Stratification in the NHS. London: NHS England, 2015.
13. Billings J, Blunt I, Steventon A, et al. Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30). BMJ Open 2012;**2**(4).
14. Billings J, Dixon J, Mijanovich T, et al. Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. Br Med J 2006;**333**:327.
15. Wennberg D, Siegel M, Darin B, et al. Combined Predictive Model - Final report & technical documentation. London: Health Dialog, King's Fund and New York University, 2006.
16. Thomson K, Lewis G. Information Governance and Risk Stratification: Advice and Options for CCGs and GPs. London: NHS England, 2013.
17. Lewis G, Curry N, Bardsley M. Choosing a Predictive Risk Model: A guide for commissioners in England. London: Nuffield Trust, 2011.
18. Berwick DM, Nolan TW, Whittington J. The Triple Aim: Care, Health, And Cost. Health Aff (Millwood) 2008;**27**(3):759-69.
19. The King's Fund. Predictive risk Project - Literature Review. London: The King's Fund, 2005.
20. Lewis G, Kirkham H, Duncan I, et al. How Health Systems Could Avert 'Triple Fail' Events That Are Harmful, Are Costly, And Result In Poor Patient Satisfaction. Health Aff (Millwood) 2013;**32**(4):669-76.
21. Stata Statistical Software: Release 14. [program]. College Station, TX: StataCorp LP, 2015.
22. IBM SPSS Statistics for Macintosh, Version 23.0 [program]. Armonk, NY: IBM Corp, 2015.
23. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). International Journal of Epidemiology 2015;**44**(3):827-36.

24. NHS Health and Social Care Information Centre. Numbers of Patients Registered at a GP Practice - July 2015. Leeds: NHS Health and Social Care Information Centre, 2015.
25. Bardsley M, Billings J, Dixon J, et al. Predicting who will use intensive social care: case finding tools based on linked health and social care data. Age Ageing 2011;**40**:265-70.
26. Kansagara D, Englander H, Salanitro A, et al. Risk prediction models for hospital readmission: A systematic review. JAMA 2011;**306**(15):1688-98.
27. Han J, Kamber M. *Data Mining: Concepts and Techniques*. 1st ed. San Diego, CA: Academic Press, 2001.
28. Armstrong JJ, Zhu M, Hirdes JP, et al. K-Means Cluster Analysis of Rehabilitation Service Users in the Home Health Care System of Ontario: Examining the Heterogeneity of a Complex Geriatric Population. Arch Phys Med Rehabil 2012;**93**(12):2198-205.
29. Coste J, Bouyer J, Fernandez H, et al. A population-based analytical approach to assessing patterns, determinants, and outcomes of health care with application to ectopic pregnancy. Med Care 2000;**38**(7):739-49.
30. Cryer PC, Saunders J, Jenkins LM, et al. Clusters within a general adult population of alcohol abstainers. International Journal of Epidemiology 2001;**30**(4):756-65.
31. Kendig H, Mealing N, Carr R, et al. Assessing patterns of home and community care service use and client profiles in Australia: a cluster analysis approach using linked data. Health & Social Care in the Community 2012;**20**(4):375-87.
32. Pud D, Ben Ami S, Cooper BA, et al. The Symptom Experience of Oncology Outpatients Has a Different Impact on Quality-of-Life Outcomes. J Pain Symptom Manage 2008;**35**(2):162-70.
33. Everitt BS, Landau S, Leese M, et al. *Cluster analysis*. 5th ed. Chichester: John Wiley & Sons, 2011.
34. StataCorp. Stata 14 Cluster Stop reference manual. College Station, TX: Stata Press, 2015.
35. Ng SK, Holden L, Sun J. Identifying comorbidity patterns of health conditions via cluster analysis of pairwise concordance statistics. Stat Med 2012;**31**(27):3393-405.
36. Chan M, Zhu MX. Investigating the health profile of Macau Chinese. J Clin Nurs 2008;**17**(11C):352-61.
37. Borglin G, Jakobsson U, Edberg Ak, et al. Older people in Sweden with various degrees of present quality of life: their health, social support, everyday activities and sense of coherence. Health & Social Care in the Community 2006;**14**(2):136-46.
38. Roland M, Abel G. Reducing emergency admissions: are we on the right track? Br Med J 2012;**345**(e6017).
39. Georghiou T, Blunt I, Steventon A, et al. Predictive risk and health care: An overview. London: Nuffield Trust, 2011.
40. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. International Journal of Epidemiology 2005;**34**(1):215-20.
41. Hao S, Jin B, Shin AY, et al. Risk prediction of emergency department revisit 30 days post discharge: a prospective study. PloS one 2014;**9**(11):e112944.
42. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. Bmj 2015;**351**.
43. Tanio C, Chen C. Innovations At Miami Practice Show Promise For Treating High-Risk Medicare Patients. Health Aff (Millwood) 2013;**32**(6):1078-82.
44. Alderwick H, Ham C, Buck D. Population health systems: Going beyond integrated care. London: The King's Fund, 2015.



Mean future care utilisation for the risk strata - High (H), Medium (M) and Low (L) - and the four high-risk clusters - 1, 2, 3 and 4.
Figure 1
190x142mm (300 x 300 DPI)



Patterns of utilisation for the four high-risk clusters – Emergency care hospitalisations (Emg), Nonemergency hospitalisations (NonE), Outpatient attendances (OP) and GP visits (GP) versus the high-risk population mean

Figure 2

190x142mm (300 x 300 DPI)

Appendix 1

Variables included in various risk scores and variables selected for our model

Variables considered in hospital admission risk studies	Number of studies out of 30 including variable in final model[1]	Variable in PARR[2]	Variable in PARR-30[3]	Variable in Combined Predictive Model[4] (*selected variables*)	Included in initial model / *included in final model after backwards elimination*
Morbidities	Medical diagnoses or comorbidity indices: 24	Cerebrovascular disease			Any diagnosis of cerebrovascular disease in 2008-2011 (in primary or secondary care)
		Chronic obstructive pulmonary disease	Chronic pulmonary disease	COPD	***Any diagnosis of COPD in 2008-2011 (in primary or secondary care)***
				Asthma (only considered in LTC counts)	Any diagnosis of Asthma in 2008-2011 (in primary or secondary care)
		Connective tissue disease/rheumatoid arthritis			***Any diagnosis of Rheumatic disease in 2008-2011 (in primary or secondary care)***
		Developmental disability			Any diagnosis of Learning disability in 2008-2011 (in primary or secondary care)
		Diabetes	Diabetes with chronic complications	*Diabetes (only considered in LTC counts)*	Any diagnosis of Diabetes in 2008-2011 (in primary or secondary care)
		Ischaemic heart disease		*CAD (only considered in LTC counts)*	***Any diagnosis of Ischaemic heart disease in 2008-2011 (in primary or secondary care)***
		Peripheral vascular disease	Peripheral vascular disease		Any diagnosis of Peripheral vascular disease in 2008-2011 (in primary or secondary care)
		Renal failure	Renal disease		***Any diagnosis of Renal disease in 2008-2011 (in primary or secondary care)***

		Sickle cell disease			
			Metastatic cancer with solid tumour	*Cancer (only considered in LTC counts)*	***Any diagnosis of Cancer in 2008-2011 (in primary or secondary care)***
			Other malignant cancer		
			Congestive heart failure	*CHF (only considered in LTC counts)*	Any diagnosis of Congestive heart failure in 2008-2011 (in primary or secondary care)
			Moderate/severe liver disease		Any diagnosis of Liver disease in 2008-2011 (in primary or secondary care)
			Other liver disease		
			Haemiplegia or paraplegia		Any diagnosis of Paraplegia in 2008-2011 (in primary or secondary care)
			Dementia		Any diagnosis of Dementia in 2008-2011 (in primary or secondary care)
				Hypertension (only considered in LTC counts)	
		Diagnostic cost groups/hierarchical condition category			
				1 LTC	***Flag if the sum of conditions listed is 0, 1 or 2 or more***
				2+ LTCS	
Mental health morbidities	Alcohol or substance use: 11	Alcohol related diagnosis		Psychoactive substance abuse	
	Mental illness: 9			Psychotic disorder	
				Inpatient admission with diagnosis of mental illness	***Any diagnosis of Mental health disorder in 2008-2011 (in primary or secondary care)***
				Depression (only as included in LTC counts)	
Prior use of medical services	Hospitalisations: 14	Previous admission for respiratory infection			
		Previous admission for a reference condition			
		Number of emergency admissions in previous 90, 180	Whether there had been a prior emergency hospital	[Combinations of] 1, 1+, 2, 2+, 3+ emergency admissions in last 30,	***Number of emergency***

		and 365 days	discharge in the past 30 days	30 to 90, 90 to 180, 180 to 365, 365 to 730 days	***admissions in 2011*** & ***Number of emergency admissions over 2008-2011***
		Average number of episodes per spell for emergency admissions		Average number of episodes per spell for emergency admissions >=3	
			Whether the current admission was an emergency admission		
		Total number of previous emergency admissions in previous three years	Number of emergency hospital discharges in the last year		
		Number of non-emergency admissions in previous 365 days			***Number of non-emergency admissions over 2008-2011***
	Emergency department visits: 4			A&E visits and investigations	
	Clinic visits or missed visits: 3	Number of different treatment specialists seen			
				[Combinations of] 1, 1-5, 2, 3+, 6-10, 11+ out-patient specialty visits in last 30, 30 to 90, 365 to 730 days	***Number of outpatient visits over 2008-2011***
	Index hospital length of stay: 4				
	Other			Polypharmacy: 1-4 unique drugs in any month (last 0 to 90 days); 5-9; 10+	***Number of GP visits over 2008-2011 (including home visits)***
Sociodemographic factors	Age: 19	Age 65-74 or age 75+	Age squared	Age band (0-4, 15-39, 40-59, 5 year age bands, 85+)	***5-year age bands*** & Over 75 flag
	Sex: 15	Sex		Gender	***Gender***
	Race/ ethnicity: 7	Ethnicity			
Social determinants of health	SES, income and employment: 5		Index of multiple deprivation band for the place of residence		***Townsend score (5 groups)***
	Insurance status: 6				

	Education: 0			
	Marital status and people in household: 4			
	Social support: 2			
	Access to care: 5			
	Discharge location: 2			
Hospital specific metrics	*Not included in review*	Observed:expected ratio for practice style sensitive admissions in ward of residence		
		Observed:expected ratio for rate of readmissions for hospitals of current admission	Hospital-specific variable	
Illness severity	Severity index: 1			
	Laboratory findings: 4			
	Other: 4			
Overall health and function	Functional status, ADL: 2			
	Self-rated health, QOL: 3			
	Cognitive impairment: 7			
	Visual/hearing impairment: 1			

1. Kansagara D, Englander H, Salanitro A, et al. Risk prediction models for hospital readmission: A systematic review. JAMA 2011;**306**(15):1688-98.
2. Billings J, Dixon J, Mijanovich T, et al. Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. Br Med J 2006;**333**:327.
3. Billings J, Blunt I, Steventon A, et al. Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30). BMJ Open 2012;**2**(4).
4. Wennberg D, Siegel M, Darin B, et al. Combined Predictive Model - Final report & technical documentation. London: Health Dialog, King's Fund and New York University, 2006.

Supplementary file

DECIDING ON THE NUMBER OF CLUSTERS - METHODOLOGICAL EXPLORATION

To segment the high-risk population, a k-means method was used. This method is efficient even for large sample sizes and produces roughly similar sized segments.[1] However, this method also require the number of clusters (k) to be specified before the analysis, rather than deducing it from the results afterwards. Therefore, a number of steps were taking to identify the optimal number of clusters for this population.

Pseudo-F Statistic

The main method for determining the number of clusters was the Pseudo-F statistic.[2] This statistic is commonly used in healthcare clustering studies,[3-7] and has been identified as one of the best criteria to determine the number of clusters.[8] It compares the between-cluster to the within-cluster sum-of-squares, and a large Pseudo-F statistic indicates distinct clusters.[9]

The k-means analysis was run for 2 to 8 clusters, and the Pseudo-F statistic was calculated for each solution (see table 1). A peak could be observed around the 3- and 4-cluster solutions.

Table 1: Pseudo-F statistics for 2- to 8-cluster solutions

2 clusters	2249
3 clusters	2745
4 clusters	2662
5 clusters	2374
6 clusters	2267
7 clusters	2131
8 clusters	2041

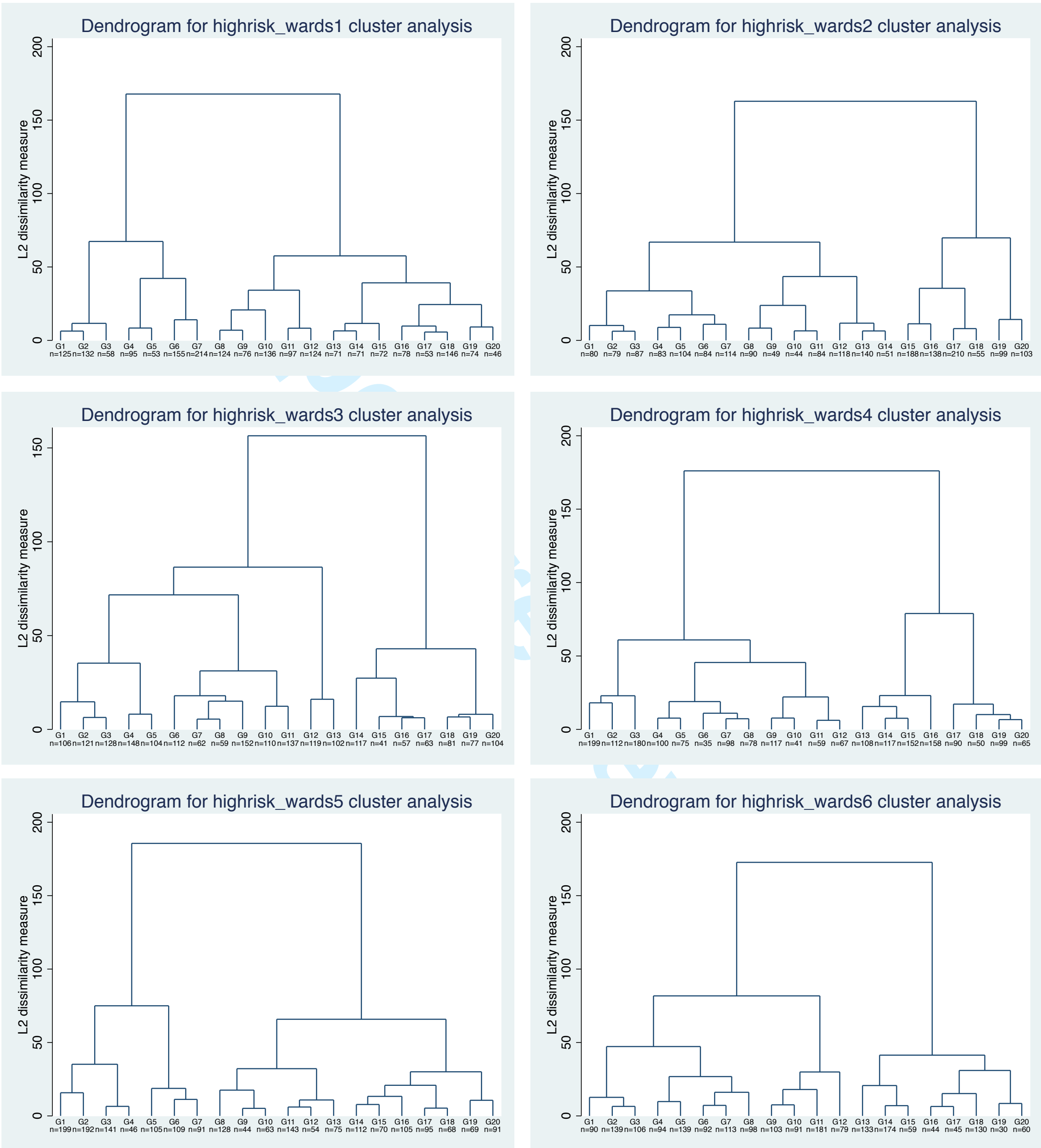
Ward's linkage

K-means is a non-hierarchical clustering method. Hierarchical methods, including the popular Ward's method, do not require k to be specified before the analysis.[8] Hierarchical clustering can be used to gain more insight into the data's structure. By displaying the results as a dendogram (a tree-like plot detailing each hierarchical step in the model) different clustering solutions can be visually explored.[10,11] Indeed, many studies combine hierarchical clustering with k-means in a two-stepped approach.[12-15]

However, hierarchical methods present some limitations. The approach is computational intensive and struggles to handle large datasets with more than a thousand observations.[10,11] In addition, hierarchical clustering based on Ward's method can be sensitive to outliers.[8]

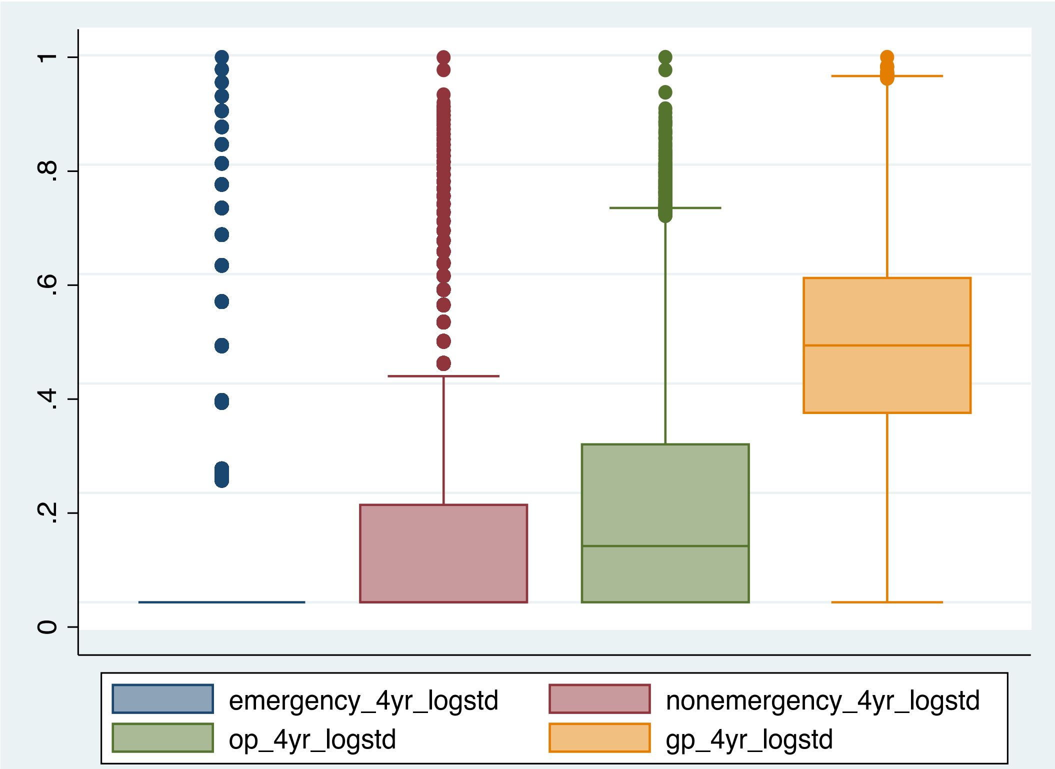
The high-risk population in the test sample, consisting of 7,433 people, was too large to include in its entirety in a hierarchical cluster analysis. Therefore, three unique, random samples of 2,000 people we used. After reshuffling the data, another three 2,000 people samples were taken and clustered. These results were then analysed through dendograms (see figure 1). All samples favoured a two-cluster solution, reflecting high- and low-utilisation groups, with the next split being further down the graph. The samples showed different results regarding the next best split. Sample 1,2, 3 and 5 can be interpreted as indicating the existence of four distinct clusters. Sample 4 favoured five clusters, and sample 6 could be interpreted as three or five clusters. Overall, the differences at this level are small.

Figure 1: *Dendograms for the six 2,000 people samples clustered using Ward's linkage*



One of the reasons the results are different across the samples is the impact of outliers, which Ward's method is sensitive to.[8] Despite the log-normalisation of the clustering variables, there still exist a large number of outliers (see figure 2). Especially in the smaller samples used for the clustering, these outliers could have changed the resulting clusters.

Figure 2: Box plots of the standardised, log-normalised clustering variables

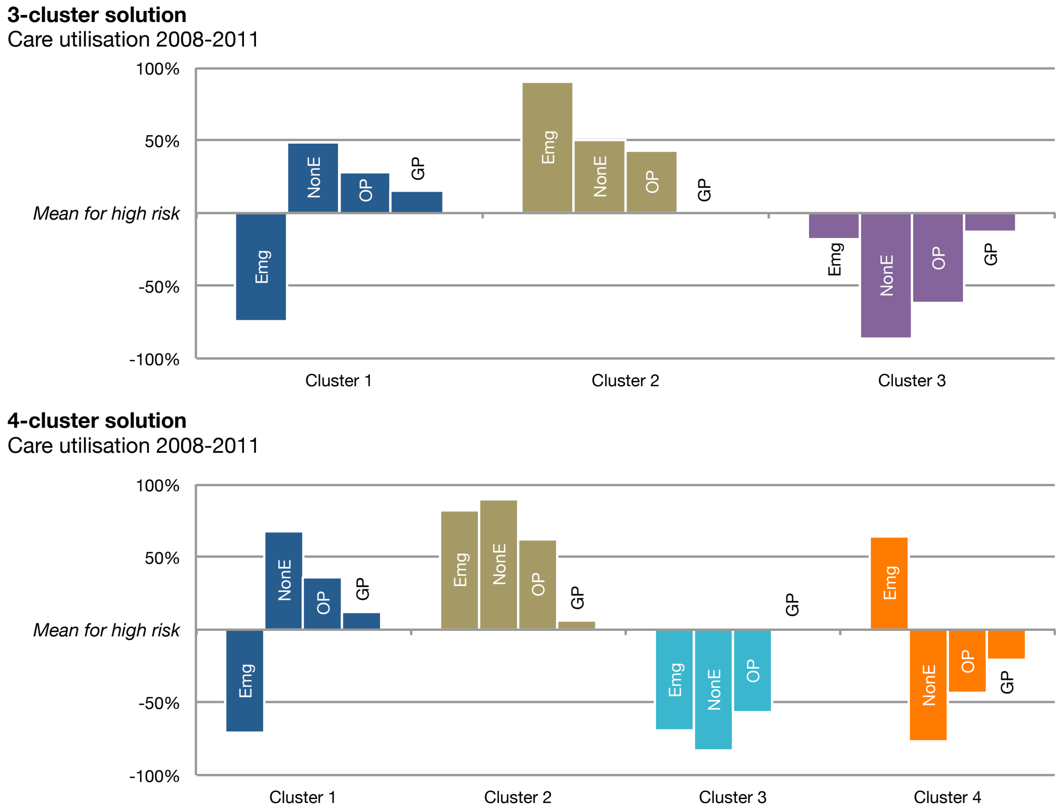


Post-hoc analysis

It is important to keep in mind that in cluster analysis, there is no absolute 'right' answer[16] - it all depends on the purpose of the clustering. Some other aspects to consider in evaluating the number of segments are, for example, interpretability, actionability and ease of use.[11]

The cluster means of the 3- and 4-cluster solutions were compared to review the practical usefulness of the resulting population groups (see figure 3). Both solutions found clusters of people with high utilisation but low emergency care use (clusters one), and people with overall high utilisations (clusters two). As the third group, the 3-cluster solution identified people with low overall utilisation but average emergency care use. However, the 4-cluster solution split this final cluster into two very distinct groups: people with overall low utilisation, and people with low utilisation but high emergency care use.

Figure 3: Practical comparison of the 3- and 4-cluster solutions



Considering the relevance of emergency care use for risk stratification, the difference between clusters three and four are important to the interpretability of the results. In terms of actionability, differentiating between these two groups allows tailored initiatives to be developed that target those people with low care utilisation but high emergency admissions. Taking this, and the previously described analyses into account, the 4-cluster solution was ultimately selected.

References

1. Han J, Kamber M. *Data Mining: Concepts and Techniques*. 1st ed. San Diego, CA: Academic Press, 2001.
2. Stata Statistical Software: Release 14. [program]. College Station, TX: StataCorp LP, 2015.
3. Armstrong JJ, Zhu M, Hirdes JP, et al. K-Means Cluster Analysis of Rehabilitation Service Users in the Home Health Care System of Ontario: Examining the Heterogeneity of a Complex Geriatric Population. Arch Phys Med Rehabil 2012;**93**(12):2198-205.
4. Coste J, Bouyer J, Fernandez H, et al. A population-based analytical approach to assessing patterns, determinants, and outcomes of health care with application to ectopic pregnancy. Med Care 2000;**38**(7):739-49.
5. Cryer PC, Saunders J, Jenkins LM, et al. Clusters within a general adult population of alcohol abstainers. International Journal of Epidemiology 2001;**30**(4):756-65.
6. Kendig H, Mealing N, Carr R, et al. Assessing patterns of home and community care service use and client profiles in Australia: a cluster analysis approach using linked data. Health & Social Care in the Community 2012;**20**(4):375-87.
7. Pud D, Ben Ami S, Cooper BA, et al. The Symptom Experience of Oncology Outpatients Has a Different Impact on Quality-of-Life Outcomes. J Pain Symptom Manage 2008;**35**(2):162-70.
8. Everitt BS, Landau S, Leese M, et al. *Cluster analysis*. 5th ed. Chichester: John Wiley & Sons, 2011.
9. StataCorp. Stata 14 Cluster Stop reference manual. College Station, TX: Stata Press, 2015.
10. IBM Corp. IBM SPSS Statistics Base 22. Chicago, IL: IBM Software Group, 2013.
11. Tsiptsis K, Chorianopoulos A. *Data Mining Techniques in CRM: Inside Customer Segmentation*. Chichester: John Wiley & Sons, 2009.
12. Lega F, Mengoni A. Profiling the different needs and expectations of patients for population-based medicine: a case study using segmentation analysis. BMC Health Serv Res 2012;**12**(1):473.
13. Liu C-Y, Liu J-S. Mining the optimal clustering of people's characteristics of health care choices. Expert Systems with Applications 2011;**38**(3):1400-04.
14. Nagel GC, Schmidt S, Strauss BM, et al. Quality of life in breast cancer patients: a cluster analytic approach. Breast Cancer Res Treat 2001;**68**(1):75-87.
15. Leijon O, Härenstam A, Waldenström K, et al. Target groups for prevention of neck/shoulder and low back disorders: an exploratory cluster analysis of working and living conditions. Work 2006;**27**(2):189.
16. Han J, Kamber M, Pei J. *Data Mining: Concepts and Techniques*. 3rd ed. Waltham, MA: Morgan Kaufmann, 2011.