BMJ Open

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

Sabine I Vuik,[1] Erik Mayer,[2] Ara Darzi[1,2]

To cite: 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;**6**: e012903. doi:10.1136/ bmjopen-2016-012903

▸ Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/ 10.1136/bmjopen-2016-012903).

Received 2 June 2016
Revised 23 September 2016
Accepted 2 November 2016

[1]Institute of Global Health Innovation, Imperial College, St Mary's Hospital, London, UK
[2]Department of Surgery, Imperial College, St Mary's Hospital, London, UK

Correspondence to
Sabine I Vuik;
s.vuik@imperial.ac.uk

ABSTRACT

Objective: To show how segmentation can enhance risk stratification tools for integrated care, by providing insight into different care usage 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 the usage of different care settings using a k-means cluster analysis. Data from 2008 to 2011 were used to create the risk score and segments, while 2012 data were used to understand the predictive abilities of the models.

Setting and participants: Data were collected from administrative data sets 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 usage of 4 different care settings: emergency acute care, elective acute care, outpatient care and GP care.

Results: While the risk strata predicted care usage at a high level, within the high-risk population, usage varied significantly. 4 different groups of high-risk patients could be identified. These 4 segments had distinct usage patterns across care settings, reflecting different levels and types of care needs. The 2008–2011 usage patterns of the 4 segments were consistent with the 2012 patterns.

Discussion: Cluster analyses revealed that the high-risk population is not homogeneous, as there exist 4 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: Usage-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 data set containing patient-level linked primary and secondary care administrative data.
- Rather than focusing only on emergency care, this study looks at patterns of usage 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 were for a random sample of English patients, and may not reflect local trends.
- No data were available in linked format for other care settings, such as accident and emergency, mental health, community and social care.

BACKGROUND

In healthcare, a small number of patients accounts for a disproportionally large share of usage.[1,2] Risk stratification can be applied to identify and target this group. 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. First, this approach only looks at one element of care. While the risk of an emergency hospitalisation can be expected to correlate with the overall use of emergency acute care, usage 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.

Second, detailed information on the characteristics of the high-risk patients, such as age, morbidities and socioeconomic 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 usage-based segmentation can enhance risk stratification tools used for integrated care by, first, taking into account care usage across multiple care settings and, second, 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 usage 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 'historical' (2008–2011) and 'future' (2012) data sets. The historical data set reflects the information that would be available to healthcare professionals conducting risk stratification and cluster analysis at the end of 2011, while the future data set was used to understand how accurately the models predicted actual usage in the following year.

Software

STATA (V.14) (Stata Statistical Software: Release 14. [program]. College Station, Texas: StataCorp LP, 2015) was used to perform the cluster analyses and calculate the pseudo-F statistics. For all other analyses, including the risk prediction, SPSS (V.23) (IBM SPSS Statistics for Macintosh, Version 23.0 [program]. Armonk, New York: IBM Corp, 2015) was used.

Data

A data set 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 data set is broadly representative of the age, sex and ethnicity composition of the UK population.[21] 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 centile,[22] to reflect a typical local population in England.

The final data set included patient demographics, long-term condition (LTC) diagnoses and usage variables. We selected four high-level usage variables for the cluster analysis of high-risk patients: inpatient emergency hospitalisations, inpatient non-emergency hospitalisations, outpatient attendances and GP visits. These usage variables were used to reflect different care settings that may be incorporated in integrated care models. For the cluster analysis, the usage 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 23] 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 24] as detailed in online supplementary appendix 1. The logistic regression on the training set excluded a number of diagnosis variables after stepwise 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 24] The test population was stratified into three groups, which comprised the top 5% highest risk patients ('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]

Table 1 Strata characteristics

	High risk	Medium risk	Low risk	Total population
Number of people	7466	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 (IQR)	2 (1–2)	1 (0–1)	0 (0–0)	0 (0–0)
Number of emergency hospitalisations over 2008–2011, median (IQR)	1 (1–3)	0 (0–1)	0 (0–0)	0 (0–0)
Number of nonemergency hospitalisations over 2008–2011, median (IQR)	1 (0–3)	1 (0–2)	0 (0–0)	0 (0–1)
Number of outpatient attendances over 2008–2011, median (IQR)	16 (8–30)	8 (2–16)	1 (0–4)	1 (0–6)
Number of GP visits over 2008–2011 median (IQR)	55 (35–82)	34 (22–51)	10 (4–20)	13 (6–27)
Number of emergency hospitalisations in 2012, median (IQR)	0 (0–1)	0 (0–0)	0 (0–0)	0 (0–0)
Number of non-emergency hospitalisations in 2012, median (IQR)	0 (0–1)	0 (0–0)	0 (0–0)	0 (0–0)
Number of outpatient attendances in 2012, median (IQR)	4 (1–8)	1 (0–4)	0 (0–1)	0 (0–2)
Number of GP visits in 2012, median (IQR)	13 (7–22)	8 (5–14)	2 (0–5)	3 (1–7)

Segmentation

For the segmentation analysis, the k-means algorithm was used to cluster the patients based on their historical usage. This method was selected as it is efficient and produces roughly similar-sized segments.[25] 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,[26–30] and is one of the best criteria to determine the number of clusters.[31] It compares the between-cluster with the within-cluster sum-of-squares, and a large pseudo-F statistic indicates distinct clusters.[32] In addition, the different clustering solutions were also explored using Ward's linkage clustering and post hoc analysis, as detailed in online supplementary appendix 2. 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 usage of each cluster to the historical pattern.

Analysis

To create profiles for the segments, the usage variables as well as demographic characteristics were analysed to see if they differed significantly across segments. For the non-normal usage 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 χ^2 test. Where these tests found significant variation across segments, the results were then explored pairwise between segments to identify which segment or segments were significantly different from others. For this, the Mann-Whitney U tests, Student's 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 data set 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, LTCs and historical care usage were all found to increase with each risk stratum (see table 1). In addition to historical usage, future usage of all care types also increased for the high-risk stratum.

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

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

The clusters demonstrate a great variation in future care usage within the high-risk stratum (see figure 1). Emergency care usage, which defines high-risk patients, is high for all clusters. Nevertheless, clusters 1 and 3 have emergency care usage rates that lie closer to the medium risk stratum than the high-risk average. Non-emergency hospitalisations and outpatient

Table 2 Clusters within the high-risk population

	Cluster				ANOVA/Kruskal-Wallis/χ^2 test
	1	2	3	4	
Clustering variables					
Number of emergency hospitalisations over 2008–2011, median (IQR)	1 (0–1)‡	3 (2–4)*	1 (0–1)‡	3 (2–4)*	KW: <0.000
Number of non-emergency hospitalisations over 2008–2011, median (IQR)	3 (2–5)*	3 (2–5)*	0 (0–1)*	0 (0–1)*	KW: <0.000
Number of outpatient attendances over 2008–2011, median (IQR)	24 (16–38)*	29 (18–46)*	7 (3–13)*	10 (5–18)*	KW: <0.000
Number of GP visits over 2008–2011, median (IQR)	61 (43–90)*	57 (40–86)*	55 (35–82)*	42 (26–65)*	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‡	χ^2: <0.000
Age at end of study period, mean	79*	67*	83*	71*	AN: <0.000
Number of long-term conditions, median (IQR)	2 (1–3)‡	2 (1–3)‡	1 (1–2)*	1 (1–2)*	KW: <0.000
Number of emergency hospitalisations in 2012, median (IQR)	2 (1–3)‡	2 (1–3)‡	1 (1–2)‡	1 (1–2)‡	KW: <0.000
Number of non-emergency hospitalisations in 2012, median (IQR)	0 (0–0)‡	0 (0–1)‡	0 (0–0)*	0 (0–1)*	KW: <0.000
Number of outpatient attendances in 2012, median (IQR)	0 (0–1)*	0 (0–1)*	0 (0–0)*	0 (0–0)*	KW: <0.000
Number of GP visits in 2012, median (IQR)	5 (2–10)*	6 (3–11)‡	2 (0–4)‡	2 (0–5)*	KW: <0.000
Prevalence of acute myocardial infarction, %	15*	23*	10*	19*	χ^2: <0.000
Prevalence of asthma, %	28†	26	24†	25	χ^2: 0.028
Prevalence of cancer, %	26*	22*	8*	5*	χ^2: <0.000
Prevalence of cerebrovascular disease, %	9‡	15‡	10‡	18‡	χ^2: <0.000
Prevalence of congestive heart failure, %	8*	13‡	5*	13‡	χ^2: <0.000
Prevalence of COPD, %	18†	17†	13*	18†	χ^2: <0.000
Prevalence of dementia, %	3‡	3‡	5‡	7‡	χ^2: <0.000
Prevalence of diabetes, %	28‡	22‡	28‡	22‡	χ^2: <0.000
Prevalence of HIV/AIDS, %	0	0	0	0	χ^2: 0.39
Prevalence of learning disabilities, %	0†	0†	0	0	χ^2: 0.032
Prevalence of liver disease, %	1	1†	0‡	1†	χ^2: <0.000
Prevalence of mental health conditions, %	2†	3†	2†	5*	χ^2: <0.000
Prevalence of paraplegia, %	1‡	3‡	1‡	3‡	χ^2: <0.000
Prevalence of peptic ulcer, %	4†	4†	2‡	3	χ^2: <0.000
Prevalence of peripheral vascular disease, %	8*	11*	4‡	6‡	χ^2: <0.000
Prevalence of renal disease, %	23†	23†	24†	18*	χ^2: <0.000
Prevalence of rheumatic disease, %	10‡	8†	6†	5‡	χ^2: <0.000

*Significantly different from three other clusters.
†Significantly different from one other clusters.
‡Significantly different from two other clusters; all at 0.05/4=0.0125 significance level (Bonferroni adjustment).

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

DISCUSSION

Principal findings

The low, medium and high risk strata broadly correlate with care usage. For all care settings, the high-risk stratum has the highest historical and future usage. 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 usage rates, characteristics and care priorities.

Comparing historical and future usage for the four clusters, similar patterns can be observed, indicating that cluster analysis of historical data can help predict future needs. However, future usage rates were closer to the group mean for all clusters and all care settings than historical rates. This can be at least partially explained by regression to the mean (RTM), which is known to affect care usage predictions.[12 36 37] RTM describes the phenomenon where exceptionally high or low observations tend to be followed by less extreme observations in repeated measurements.[38] This effect is compounded if participants are stratified based on baseline measurements, which is the case when patients are clustered based on their 2008–2011 usage.

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 39 40] or care for,[6 8 10 11 36 41] 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 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 usage of non-acute care settings as predictor variables,[15] this detail is lost in the final risk score and the stratification. A usage-based segmentation analysis, as demonstrated in this study, can be used to bring out this detail.

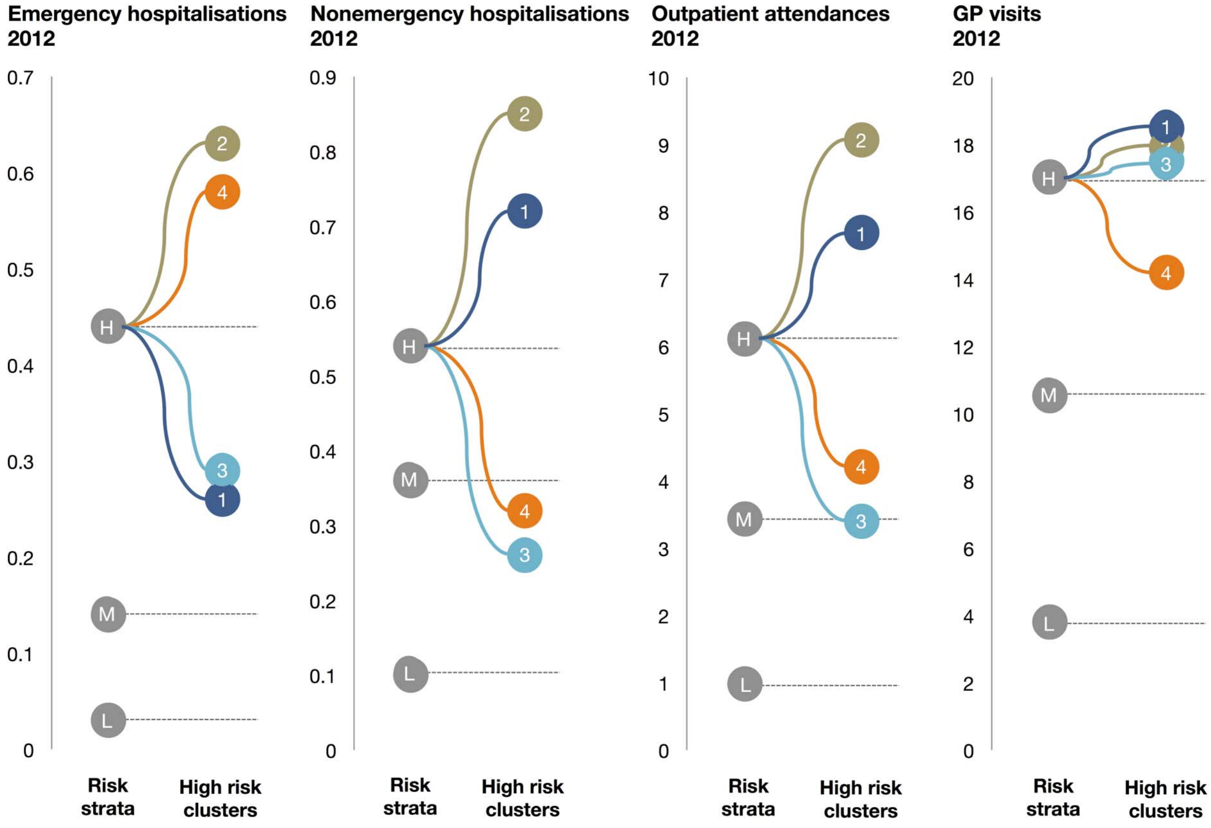

Figure 1 Mean future care usage for the risk strata—high (H), medium (M) and low (L)—and the four high-risk clusters—1, 2, 3 and 4.

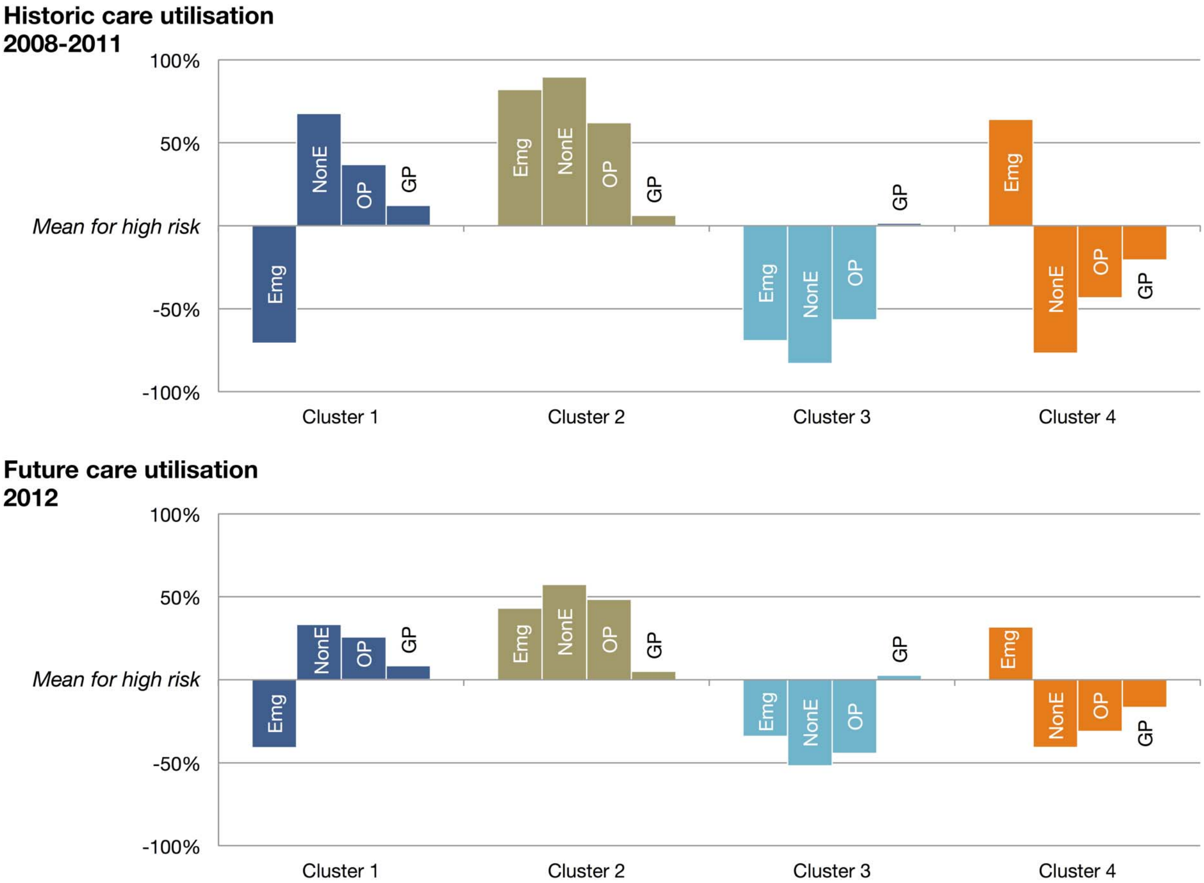

Figure 2 Patterns of usage for the four high-risk clusters—emergency care hospitalisations (Emg), non-emergency hospitalisations (NonE), outpatient attendances (OP) and general practitioner visits (GP) versus the high-risk population mean.

Limitations and future research

While 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 were available on the usage 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 performed using more extensive data sets 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 LTC 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 usage of non-emergency 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 usage rates, the highest risk score and the most LTCs. Surprisingly, this segment is also the youngest of the 4, with an average age of 67. Overall, high care usage 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 36 42] this population group does not necessarily have the highest care usage.

Cluster 4 has one of the highest usage 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 preventive 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 enhanced 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 usage 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.

Twitter Follow Sabine Vuik @sabinevuik

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

Contributors SIV designed the study, created the database, analysed the data and drafted and revised the paper. She is the 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.

Funding This study was partially funded by the Sowerby eHealth Forum, sponsored by the Peter Sowerby Foundation.

Disclaimer 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.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Technical appendix available in online supplementary files and statistical code are available from the corresponding author.

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:e007771.
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:e010301.
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:e130.
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:h6817.
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:e046.
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:pii: e001667.
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. *BMJ* 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;27:759–69.
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;32:669–76.
21. Herrett E, Gallagher AM, Bhaskaran K, *et al*. Data resource profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015;44:827–36.
22. 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.
23. 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.
24. Kansagara D, Englander H, Salanitro A, *et al*. Risk prediction models for hospital readmission: a systematic review. *JAMA* 2011;306:1688–98.
25. Han J, Kamber M. *Data mining: concepts and techniques*. 1st edn San Diego, CA: Academic Press, 2001.
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;93:2198–205.
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;38:739–49.

28. Cryer PC, Saunders J, Jenkins LM, *et al*. Clusters within a general adult population of alcohol abstainers. *Int J Epidemiol* 2001;30:756–65.
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 Soc Care Community* 2012;20:375–87.
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;35:162–70.
31. Everitt BS, Landau S, Leese M, *et al*. *Cluster analysis*. 5th edn. 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;31:3393–405.
34. Chan MF, Zhu MX. Investigating the health profile of Macau Chinese. *J Clin Nurs* 2008;17:352–61.
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 Soc Care Community* 2006;14:136–46.
36. Roland M, Abel G. Reducing emergency admissions: are we on the right track? *BMJ* 2012;345:e6017.
37. Georghiou T, Blunt I, Steventon A, *et al*. *Predictive risk and health care: an overview*. London: Nuffield Trust, 2011.
38. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. *Int J Epidemiol* 2005;34:215–20.
39. 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:e112944.
40. 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:h3868.
41. Tanio C, Chen C. Innovations at Miami practice show promise for treating high-risk medicare patients. *Health Aff (Millwood)* 2013;32:1078–82.
42. Alderwick H, Ham C, Buck D. *Population health systems: going beyond integrated care*. London: The King's Fund, 2015.