

Identifying high-risk patient clusters for falls using unsupervised machine learning on linked primary and secondary care electronic health record data

Received: 15 January 2026

Accepted: 25 May 2026

Published online: 06 July 2026

Cite this article as: Lear R., Tao R., Carenzo C. *et al.* Identifying high-risk patient clusters for falls using unsupervised machine learning on linked primary and secondary care electronic health record data. *Sci Rep* (2026). <https://doi.org/10.1038/s41598-026-55527-4>

Rachael Lear, Rachel Tao, Catalina Carenzo, Ekin Yagis, Zhuoyu Li, Maria Tipper, Agnes Motsi, Sharon O'Malley, Yusuf Abdullahi, Ben Glampson, Colin Mitchell & Erik K. Mayer

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Identifying High-Risk Patient Clusters for Falls Using Unsupervised Machine Learning on Linked Primary and Secondary Care Electronic Health Record Data

Rachael Lear^{1,4}, Rachel Tao^{1,4}, Catalina Carenzo^{1,4}, Ekin Yagis^{1,4,*}, Zhuoyu Li^{1,4}, Maria Tipper³, Agnes Motsi³, Sharon O'Malley³, Yusuf Abdullahi^{1,4}, Ben Glampson^{1,3,4}, Colin Mitchell³, Erik K Mayer^{1,2,3,4}

¹ iCARE Secure Data Environment & Digital Collaboration Space, NIHR Imperial Biomedical Research Centre, London, UK

² NIHR North West London Patient Safety Research Collaboration, Institute of Global Health Innovation, Imperial College London, UK

³ Imperial College Healthcare NHS Trust, UK

⁴ Faculty of Medicine, Department of Surgery & Cancer, Imperial College London, UK

*eyagis@ic.ac.uk

ABSTRACT

Falls risk is multifactorial, involving a combination of clinical and sociodemographic factors. Although guidelines acknowledge this complexity, most research has focused on individual risk factors, leaving the combined impact of comorbidities relatively understudied. This population-wide study used electronic health records (EHR) linked across primary and secondary care to identify falls risk profiles in the North West London (NWL) population and to stratify patients by their likelihood of requiring falls-related hospital care using unsupervised clustering. We conducted cluster analysis on patients from NWL General Practice records using coded falls risk factors. Cluster membership was compared against the risk of falls-related hospital encounters. Among four identified clusters, two groups of older, multimorbid patients were 11 times more likely to have a fall-related hospital encounter (RR 11.45, 95% CI 10.14–12.92 and RR 11.63, 95% CI 10.30–13.13) and had significantly longer mean length of stay compared with younger, fitter patients. Between two younger clusters, patients with higher deprivation levels were 29% more likely to have a fall-related hospital encounter (RR 1.29, 95% CI 1.12–1.49). These findings demonstrate that clustering routinely collected EHR data can identify population segments at highest risk of falls-related hospital use, supporting more targeted, multifactorial risk assessment and prevention strategies.

Index Terms—Falls, Risk Factors, Population Health, Machine Learning, Electronic Health Records

Introduction

Despite efforts by carers and health and social care services to manage falls risk in community and healthcare settings, around 1 in 3 adults over 65 years fall each year and falls in younger people are increasing [1], [2], [3]. The consequences can be devastating, not only due to physical injuries but also the fear of future falls, loss of confidence and independence, and the associated impact on family and caregivers [1], [2], [4]. In the English National Health Service (NHS) falls account for more than 4 million hospital bed days with annual expenditure on injurious falls over £2 billion [5]. Ageing populations and rising rates of multimorbidity, polypharmacy, and frailty, mean that falls and falls-related injuries are expected to increase

further [1], [4], [5], [6]. Furthermore, falls have consistently ranked at the top of reported patients' safety incident reports for several years, highlighting this as an unresolved issue that necessitates a new approach to prevention and management [7], [8]. As such, better targeted preventative measures will be key to reducing the harm and health and social care costs [9], [10].

Falls are usually the result of multiple contributory factors, which may be intrinsic (patient-related), extrinsic (environment-related), and behavioural (activity-related) [11], [12], [13]. Most interventions to prevent falls address individual patient risk factors, however, population-level interventions can be designed to address falls risk in whole communities, for example, implementation of policies on vitamin D supplementation and public health programmes promoting physical exercise [2], [14], [15]. Few studies have examined how specific comorbidity combinations influence hospital resource use for falls, leaving this an important yet understudied area beyond general multimorbidity and frailty. Determining multifactorial risk from diverse clinical and sociodemographic factors is important for understanding patterns of falls risk and identifying patient groups most likely to benefit from tailored prevention and monitoring strategies, however the effectiveness of population-based strategies in falls prevention remains uncertain due to the historic challenges of synthesising evidence across populations, particularly where population characteristics are poorly reported [15]. The maturity of Electronic Health Records (EHRs), clinical coding and healthcare data linked across multiple healthcare organisations now offers the possibility to tackle falls risk using a preventative health management approach, whereby population-level analysis could identify priority cohorts and risk stratified groups for targeted interventions [16].

However, few studies have used unsupervised machine learning to comprehensively explore falls risk using routinely collected EHR data at a population-wide scale. In our study, we leveraged k-means clustering to identify distinct patterns of falls risk factors, allowing for more targeted and clinically actionable insights into individual-level risk. We defined falls-related hospital encounters as ED presenting falls, ED fall diagnoses, or fall diagnoses recorded in the first inpatient episode.

By identifying clinically meaningful patient clusters using routinely collected EHR data, this study provides a foundation for developing translational, real-world tools that enable personalized falls risk assessment to reduce the number of falls and fall-related serious harm incidents.

Related work

There has been a growing interest in using machine learning to better understand and predict falls risk. These efforts aim to uncover key predictors and build models that can support earlier, more accurate risk detection in both hospital and community settings. Some prior research in this area has used EHR data, while others leverage wearable sensor inputs to model risk. Though sensor-based approaches offer real-time monitoring, EHR data allows for scalable, longitudinal analysis across diverse patient populations and is more readily generalisable across healthcare settings and EHR systems than sensor-based approaches. For instance, Archer et al. [17] recently introduced the eFalls tool—a multivariable prediction model built on EHR data for forecasting fall or fracture-related Emergency Department (ED) attendance or hospitalization in older adults. However, their approach was limited to supervised learning and relied on pre-labeled outcomes, which are labor-intensive and may overlook latent heterogeneity in risk profiles.

Unsupervised ML techniques - particularly clustering algorithms - present a promising alternative for uncovering hidden structures in health data without relying on predefined labels. Finch et al. [18] applied cluster analysis to hospital admission records to identify priority population segments for falls prevention, but this analysis lacked integration with

primary care data. Similarly, Court-Brown et al. [19] investigated socio-demographic associations with fall-related fractures but did not examine how combinations of comorbidities influenced healthcare utilization.

To date, very few studies have employed unsupervised clustering techniques on linked EHR data to identify distinct risk phenotypes. No known work has systematically combined dimensionality reduction (e.g., Principal Component Analysis) with clustering over a large, integrated dataset spanning general practice, hospital, and social care records. Our work fills this gap by leveraging the rich, multi-source Whole Systems Integrated Care (WSIC) dataset from North West London (NWL) to identify discrete patient clusters with shared falls risk characteristics and quantify their association with hospital encounters. Importantly, while multimorbidity and frailty are known fall predictors [4], [20], prior studies have not examined how specific combinations of risk factors disproportionately drive healthcare burden, nor have they explored variations by ethnicity, or deprivation status in such detail. By identifying high-risk, high-burden patient clusters through unsupervised learning, this study contributes a scalable, data-driven framework to guide precision targeting of preventive interventions.

Methods

Data source and case selection criteria

The Whole Systems Integrated Care (WSIC) database holds de-identified, routinely collected EHR data for people who are registered with a NWL GP or live in NWL area, covering a population of 2.8 million [21]. Data from over 360 GPs, 10 acute and specialist hospitals, 8 social care organisations, 2 mental health trusts and 2 community health trusts are linked at patient level and held in the ICHT Secure Data Environment (SDE) within NHS infrastructure [21]. Our study population included all patients ≥ 18 years old who were registered with a NWL GP between 1st June 2021 and 31st May 2023 (Figure S1). Using clinical coding (ICD-10, SNOMED, and GP Read), we identified i) patients with 1 or more falls risk factors and ii) patients with a falls-related hospital encounter between 1st June 2022 – 31st May 2023. The time period for inclusion in the study extends 1 year further into the past than our criteria for patients who were hospitalized for a fall to ensure that relevant risk factors coded more than one year prior to a fall are included. To define a falls-related hospital encounter, we considered patients who had a falls code listed as either the Emergency Department (ED) reason for attendance, or any ED diagnosis, or any diagnosis during the first episode of an inpatient hospital spell. Sociodemographic data included age, sex, ethnicity, primary language, index of multiple deprivation (IMD) decile, residential location (London Borough). Deprivation was measured by IMD decile (1=most deprived, 10=least deprived); “high deprivation” refers to deciles 1-5; “low deprivation” to deciles 6-10. All methods were performed in accordance with the relevant approvals provided by the NW London Sub-Data Research Access Group (SDRAG Ref ID 342) to access the Health Research Authority approved research database (reference 18/WM/0323; IRAS project ID 252449).

Falls risk factors and definitions

Where possible, published code lists from the Quality and Outcomes Framework (QOF, business rules v47.0 [22]), the electronic Frailty Index (eFI) [23], and Healthcare Data Research UK’s (HDR UK) Phenotype Library [24] were used to identify falls risk factors (Table S1). For conditions where no previously validated code list was available, a nurse researcher (RL) collated relevant codes using SNOMED CT browser with input from a geriatric consultant (CM) and a clinical lead for falls prevention (MT) to ensure code lists were comprehensive. Multi-morbidity was defined as the presence of at least two coded long-term conditions at the start of the evaluation period, with long-term conditions defined using QOF. Polypharmacy was defined as prescription of ≥ 5 medications daily while hyper-polypharmacy was defined as prescription of ≥ 10 medications daily [25]. Prescription of high-risk medications for falls was defined using the National Falls Prevention Coordination Group’s Falls Risk Increasing Drugs List (2023) [26].

Statistical analysis

The sociodemographic and falls risk factor profile of the cohort was summarised using descriptive statistics: counts and proportions for individual risk factors and Pearson

correlations between risk factor categories. A cluster analysis was used to identify homogenous groups of sociodemographic and falls risk factor profiles and evaluate how these clusters differ in terms of healthcare utilization after a fall (ED attendance and admission to hospital, diagnostic imaging, hospital length of stay (LOS)). Principal Components Analysis (PCA) using two components combined with k-means clustering was determined the most suitable method for extracting meaningful insights from our dataset. *Two components were selected on the basis that they accounted for 80% of the total dataset variance (Figure S2). OR PCA was selected over nonlinear dimensionality reduction methods such as t-SNE and UMAP due to its interpretability, stability, and suitability for downstream clustering. We acknowledge that nonlinear interactions between comorbidities may not be fully captured by PCA; techniques such as UMAP could be explored in future work.*

1) Dimensionality reduction

PCA was used to reduce the dimensionality of the dataset [27]. The proportion of the dataset's variance that is explained by each component was assessed using a Cumulative Explained Variance plot. The first two components represented 80% of the variance (Figure S2).

2) Cluster identification

Clusters were identified using the k-means algorithm [28]. The elbow method was applied by plotting within-cluster sum of squares (inertia) across $k = 2-10$; the inflection point at $k = 4$ was selected as the optimal solution (Figure S3). Post hoc analyses with Bonferroni adjustments ($p < 0.0083$ considered statistically significant) were performed to compare cluster proportions for each fall-related factor (Table S2). To validate cluster stability, silhouette analysis was performed on a random subsample of 10,000 observations (selected for computational tractability given the cohort size), yielding a silhouette score of 0.354 for the four-cluster solution, confirming adequate separation. Comparison of alternative clustering algorithms (e.g., GMM, hierarchical clustering) was not performed and is acknowledged as a future direction.

3) Outcome evaluation

Risk ratios (RR) were calculated to estimate the relative risk of requiring falls-related hospital care based on cluster membership amongst those with a falls risk factor, using cluster 1 as the reference group for comparison.

A sub-group of patients was identified with a chief complaint or primary diagnosis code for falls in the ED, or a primary or secondary diagnosis code for falls at their first episode of their Hospital Admission. We calculated RR to estimate the probability of receiving diagnostic imaging (Computerised axial Tomography (CT scan), magnetic resonance imaging (MRI), plain radiography (x-ray)) for each cluster. For patients admitted to hospital as inpatients, a one-way ANOVA was performed with cluster membership as the independent variable and hospital LOS as the dependent variable.

SQL Server Management Studio was used to pull together our cohort from the WSIC database, select relevant clinical codes, and harmonise the data. Clustering analysis was performed using Python version 3.7.9. Outcome evaluation was performed using R version 4.2.1.

Data access and approvals

This study was undertaken within a research database that was given favourable ethics approval by the West Midlands Solihull Research Ethics Committee (reference 18/WM/0323; IRAS project ID 252449). This study was approved by the NW London Sub-Data Research Access Group (SDRAG Ref ID 342). All data used in this work were fully anonymised before analysis.

Results

Sociodemographic characteristics

Between 1st June 2021 and 31st May 2023, 1,648,834 adult patients were registered with a GP in NWL. Around three quarters of the population (75.7%) were aged between 18 and 59, just under half (46.9%) were female, and one quarter of patients (24.9%) did not have English as their first language. The population was diverse in terms of ethnic and socioeconomic background (Table S3).

Falls risk factor profile

Among the 1,648,834 adult patients, just over half (51.85%, n=854,914) had one or more risk factors for falls coded in their medical records; 15.29% had 5+ documented falls risk factors (Table 1). Around 1 in 4 patients were older than 60 years (24.30%, n=401,301) and 10.35% (n=170,685) had a previous history of falls. The most common falls risk factors were multi-morbidity (29.52%, n=486,760), frailty (20.55%, n=338,786) and polypharmacy (13.19%, n=217,456). Prevalent diseases known to contribute to falls were anaemia (13.54%, n=223,248), depression (12.58%, n=207,436), arthritis (9.36%, n=154,273) and diabetes (9.22%, n=152,073). Mobility issues, reduced functional ability, and cognitive impairment were not commonly coded in EHR (4.76%, 3.85%, and 3.47%, respectively) despite the high prevalence of frailty (20.55%, n=338,786). The risk factor categories most highly correlated with one another were polypharmacy and history of falls ($r = 0.28$). Polypharmacy ($r = 0.27$) and regular prescription of falls risk inducing drugs ($r = 0.21$) were correlated with multimorbidity and were among the highest of all assessed risk factor category correlations (Figure S4). Among the comorbidities identified, conditions such as cardiovascular disease, frailty, and polypharmacy were consistent with known falls risk factors. Diabetes mellitus also emerged as a contributing feature - likely reflecting indirect associations through neuropathy, frailty, or multimorbidity, consistent with patterns reported in recent population-level studies.

Table 1. Prevalence of falls risk factors among adults registered with a NWL GP.

	Sub-category	Count	%
History of falls			
	Falls history ^(a)	170,658	10.35%
Mobility Problems			
	Foot or lower limb problems	54,143	3.28%
	Mobility issue - general	20,087	1.22%
	Transfer problems	12,313	0.75%
	Problems with gait	8,864	0.54%
	Use of walking aid	5,158	0.31%
	Fear of falling	3,219	0.20%
	Muscle weakness	779	0.05%
	<i>Mobility Problems - All</i>	<i>78,463</i>	<i>4.76%</i>
Sensory Problems			
	Vision problem ^(b)	82,417	5.00%
	Hearing problem	39,363	2.39%
	Dizziness - vestibular issue	34,134	2.07%
	<i>Sensory Problems - All</i>	<i>133,157</i>	<i>8.08%</i>
Reduced functional ability			
	Housebound	44,462	2.07%
	Reduced functional ability	28,123	1.71%
	Activity limitation	604	0.04%
	<i>Reduced functional ability - All</i>	<i>63,446</i>	<i>3.85%</i>
Impaired cognition			
	Memory & cognitive problems ^(c)	57,260	3.47%
Learning Disability			
	Learning Disability	14,067	0.85%
Medication			
	Polypharmacy ^(d)	217,456	13.19%
	Hyper-polypharmacy ^(e)	74,606	4.52%
	Falls-Risk Increasing Drug ^(f)	134,617	8.16%
Urinary incontinence			
	Urinary incontinence	122,629	7.44%
Poor nutritional status			

	Obesity	197,948	12.01%
	Vitamin D deficiency	73,941	4.48%
	Excessive alcohol intake	6,822	0.41%
	Malnutrition	263	0.02%
Cardiovascular disorder			
	Hypotension & syncope	75,798	4.60%
	Brady- or tachyarrhythmia	31,442	1.91%
	<i>Cardiovascular disorders - All</i>	<i>97,173</i>	<i>5.89%</i>
Other contributing disease			
	Anaemia	223,248	13.54%
	Arthritis	154,273	9.36%
	Depression	207,436	12.58%
	Diabetes	152,073	9.22%
	Electrolyte disorder	24,500	1.49%
	Frailty	338,786	20.55%
	Multiple Sclerosis	2,933	0.18%
	Osteoporosis	6,428	0.39%
	Parkinson's Disease	3,412	0.21%
	Peripheral neuropathy	5,673	0.34%
	Stroke & TIA	28,272	1.71%
	Patients with at least 1 disease known to contribute to falls ^(g)	620,195	37.61%
Multi-morbidity			
	Multi-morbidity ^(h)	486,760	29.52%
Falls risk factor count			
	1 falls risk factor	257,846	15.64%
	2 falls risk factors	153,052	9.28%
	3 falls risk factors	106,639	6.47%
	4 falls risk factors	81,051	4.92%
	5 falls risk factors	256,326	15.55%
<i>Population with 1+ risk factors</i>	<i>Patients with at least 1 risk factor for falls</i>	<i>854,914</i>	<i>51.85%</i>
Total	Adult patients registered with a GP in North-west London	1,648,834	100.00%

^(a) Patients with 1 or more fall(s) recorded in their medical records *any time* before 01.06.2023.

^(b) Includes eye conditions known to affect vision; 'wears glasses' excluded due to coding quality.

^(c) Includes dementias, confusion, delirium, and behaviours that challenge.

^(d) Patients prescribed ≥ 5 (regular) medications any time 01.06.21 – 31.05.23.

^(e) Patients prescribed ≥ 10 (regular) medications any time 01.06.21 – 31.05.23.

^(f) Patients prescribed a 'high-risk' falls risk increasing drug (regular or acute prescription) any time 01.06.21 – 31.05.23. The National Falls Prevention Coordination Group (2023) considers these 'high-risk' medications to have high potential to cause a fall.

^(g) Patients with at least 1 disease known to contribute to falls recorded in their medical records *any time* before 01.06.2023.

^(h) Multi-morbidity: ≥ 2 Long Term Conditions recorded *any time* before 01.06.2023. Long Term Conditions considered: asthma, atrial fibrillation, cancer, chronic kidney disease, chronic obstructive pulmonary disease, dementia, depression, diabetes, epilepsy, heart failure, hypertension, learning disability, mental health disorder, non-diabetic hyperglycaemia, osteoporosis, peripheral arterial disease, palliative care, rheumatoid arthritis, coronary heart disease, stroke and transient ischaemic attack.

Risk factor clusters

The cluster analysis of 854,914 patients with 1+ falls risk factors identified four clusters of risk factors for falls (Tables 2 & 3). There was little variation in the ratio of males to females between the clusters; within clusters, females were over-represented compared to males. Obesity and multimorbidity were highly prevalent in all clusters ($>19\%$ and $>40\%$, respectively) and more than 10% of patients in each cluster were prescribed a Falls Risk

Increasing Drug (FRID).

Cluster 1: low age + low risk factor count + low deprivation (n= 168,627)

In this cluster, mean age was 43.0 (SD 12.8) and mean risk factor count was 2.55. Nearly half of the patients were Asian or British Asian (47.13%). Deprivation levels were low (all scores ≥ 6). Mobility issues, sensory issues, and cognitive impairment were coded for less than 5% of patients. The prevalence of anaemia was highest in this cluster (28.17%) and vitamin D deficiency was also common (10.45%).

Cluster 2: low age + low risk factor count + high deprivation (n= 313,056)

Cluster 2 was similar to cluster 1 in terms of age (mean age 41.7 (SD 12.4)), ethnic profile and mean risk factor count (2.69); however, deprivation levels were high in this group (all scores 1-5). Like cluster 1, the prevalence of mobility issues, sensory issues, and cognitive impairment were low. Depression was highest in this cluster (26.29%).

Cluster 3: high age + high risk factor count + low deprivation (n= 156,888)

Older patients were over-represented in this cluster - more than half were over the age of 70 (mean age 71.0 (SD 11.6)). Multi-morbidity and polypharmacy were highly prevalent (71.52% and 25.83%, respectively) and the mean risk factor count was 4.69. Deprivation levels in this cluster were low (all scores > 6). This cluster had the highest prevalence of cardiovascular disorders and the highest proportion of patients taking FRIDs.

Cluster 4: high age + high risk factor count + high deprivation (n= 216,343)

Mean risk factor count for this cluster was 5.12. Like cluster 3, patients in this cluster were older (mean age 68.7 (SD 11.6)); almost all patients (99.07%) were > 50 years old and 61.04% were White. Deprivation levels were high (all scores ≤ 5). This cluster has the highest prevalence of multi-morbidity (76.69%) and hyper-polypharmacy (17.37%) as well as the highest proportions of many falls risk factors including cognitive impairment, mobility problems, sensory issues, urinary incontinence, and others; the highest prevalence of severe frailty was in this cluster and nearly a third of patients (32.70%) had a history of falls.

Table 2. Description of sociodemographic characteristics among clusters

Legend: percentage (%) of patients within a cluster

<5%
5-9%
10-19%
20-29%
30-39%
40-59%
$\geq 60\%$

		Cluster 1: n= 168,627 19.72%	Cluster 2 n= 313,056 36.62%	Cluster 3 4 n= 156,888 18.35%	Cluster 4 n= 216,343 25.31%
<i>Age</i>					
	18-30	16.74	18.63	0.00	0.00
	30-39	23.87	25.50	0.00	0.00
	40-49	30.51	29.78	0.05	0.93
	50-59	16.14	17.16	20.01	23.77
	60-69	12.68	8.92	23.59	28.45
	70-79	0.06	0.03	31.66	27.51
	80-89	0.00	0.00	19.05	15.21
	90+	0.00	0.00	5.64	4.12
<i>Gender</i>					

	Female	57.57	57.15	53.76	54.00
	Male	42.42	42.84	46.24	46.00
<i>Ethnicity</i>					
	Asian or Asian British	47.13	43.18	13.84	13.16
	Black or Black British	7.76	16.33	2.66	11.30
	Mixed	3.42	4.40	1.63	2.47
	Other ethnic group	10.20	9.66	8.39	12.04
	White	31.49	26.43	73.48	61.04
<i>Deprivation decile</i>					
	1	0.00	3.70	0.00	3.93
	2	0.00	16.10	0.00	17.19
	3	0.00	23.70	0.00	22.60
	4	0.00	33.19	0.00	31.68
	5	0.00	23.30	0.00	24.59
	6	37.09	0.00	28.90	0.00
	7	23.77	0.00	22.70	0.00
	8	19.83	0.00	22.87	0.00
	9	13.09	0.00	17.59	0.00
	10	6.23	0.00	8.57	0.00
<i>Primary language</i>					
	English	40.24	37.23	46.32	45.69
	Not English	24.30	31.94	12.20	20.06
	Not recorded	35.36	30.92	41.48	34.25

Table 3. Description of falls risk factors among clusters*Legend*

<1%
1-4%
5-9%
10-19%
20-29%
30-39%
40-69%
≥70%

		Cluster 1	Cluster 2	Cluster 3	Cluster 4
		Younger, low deprivation	Younger, high deprivation	Older, low deprivation	Older, high deprivation
	Mean risk factor count	2.56	2.70	4.70	5.12
	Falls history	10.03	11.33	30.32	32.70
	Multi-morbidity	40.94	44.60	71.52	76.69

Mobility Problems					
	Foot or lower limb problems	5.32	5.13	7.57	7.97
	Mobility issue - general	0.35	0.54	4.04	5.30
	Transfer problems	0.36	0.46	2.36	3.04
	Problems with gait	0.37	0.43	1.63	2.00
	Use of walking aid	0.07	0.08	1.16	1.38
	Fear of falling	0.08	0.09	0.61	0.84
	Muscle weakness	0.03	0.04	0.12	0.19
Sensory Problems					
	Vision problem	4.77	5.42	14.33	16.13
	Hearing problem	1.88	2.01	7.98	8.02
	Dizziness - vestibular issue	3.39	3.76	4.31	4.58
Reduced functional ability					
	Housebound	3.00	2.67	8.32	8.32
	Reduced functional ability	1.24	1.81	4.25	6.33
	Activity limitation	0.02	0.02	0.13	0.14
Impaired cognition					
	Memory & cognitive problems	3.02	3.45	10.44	11.55
	Learning Disability	2.26	2.42	0.49	0.88
Medication					
	Polypharmacy	8.99	9.65	25.83	26.32
	Hyper-polypharmacy	3.07	3.98	12.36	17.37
	Falls-Risk Increasing Drug	11.50	10.48	22.86	21.52
Urinary incontinence					
	Urinary incontinence	9.84	10.00	19.50	20.41
Poor nutritional status					
	Obesity	19.01	23.63	20.61	27.54
	Vitamin D deficiency	10.45	11.45	4.51	6.20
	Excessive alcohol intake	0.75	0.91	0.58	0.83
	Malnutrition	0.01	0.01	0.04	0.06
Cardiovascular disorder					
	Hypotension & syncope	12.86	11.57	16.97	16.25
	Brady-tachyarrhythmia	2.13	2.02	6.11	5.52
Other contributing disease					

Anaemia	28.17	26.69	24.02	25.20
Arthritis	8.17	8.32	30.25	30.97
Depression	24.73	26.29	19.55	24.39
Diabetes	12.22	13.83	20.19	26.11
Electrolyte disorder	0.90	1.04	5.06	5.45
Frailty - mild	17.28	17.09	24.59	24.55
Frailty - moderate	5.33	5.46	18.80	19.19
Frailty - severe	1.17	1.34	15.71	16.89
Multiple Sclerosis	0.35	0.28	0.43	0.37
Osteoporosis	0.09	0.06	2.04	1.34
Parkinson's Disease	0.08	0.06	0.96	0.73
Peripheral neuropathy	0.35	0.38	1.03	1.05
Stroke & TIA	0.92	0.95	6.29	6.42

Association with falls-related hospital encounters

During the 12-month period between 1st June 2022 and 31st May 2023, 8,095 adult patients (0.95% of the population with 1+ risk factor(s)) had a falls-related hospital encounter (i.e. ED attendance and/or inpatient admission) coded in their medical records. Most patients (99.81%, n=8,080/8,095) attended hospital via the ED and 7,633 were admitted as hospital inpatients. The risk of a falls-related hospital encounter amongst those with a falls risk factor was 11 times higher in clusters 3 (older, low deprivation level) and 4 (older, high deprivation level) (RR 11.63, 95%CI 10.30-13.13 and RR 11.45, 95%CI 10.14-12.92, respectively) compared to cluster 1 (younger, low deprivation level) (Figure 1). The risk for cluster 2 (younger, high deprivation level) was 29% higher than for cluster 1 (younger, low deprivation level), on average (RR 1.29, 95%CI 1.12-1.49). Risk ratios are presented unadjusted, as cluster membership encodes the combined sociodemographic and clinical profile of each group. The Cluster 1 vs. Cluster 2 comparison is of particular note as these clusters are age-matched, and the observed difference in falls-related hospital risk is attributable to deprivation rather than age. For those who had a falls-related hospital encounter, the probability of receiving diagnostic imaging did not differ significantly across clusters (Figure S5, Table S4), however, for falls-related hospital admissions, LOS varied significantly between clusters ($F=71.17$, $p<.0001$); mean LOS was highest in cluster 4 (mean 8.10 days, 95% CI 7.28-8.42) and lowest in cluster 2 (mean 2.98 days, 95% CI 2.54-3.42) (Figure 2, Table S4).

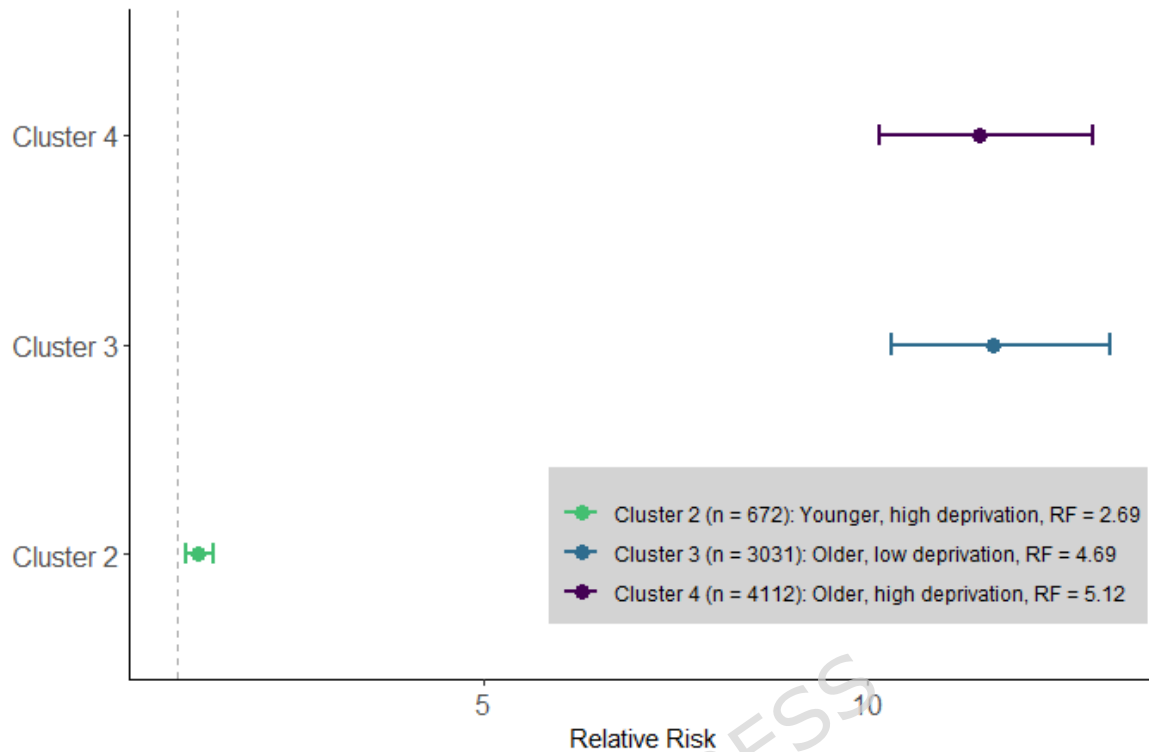


Figure 1. Relative risk of falls-related hospital encounter for Clusters 2, 3, and 4, using Cluster 1 (younger, low deprivation) as the reference group. 'RF' denotes the average number of risk factors per person in that cluster. For example, the mean number of risk factors per person in Cluster 2 (younger, high deprivation) is 2.69.

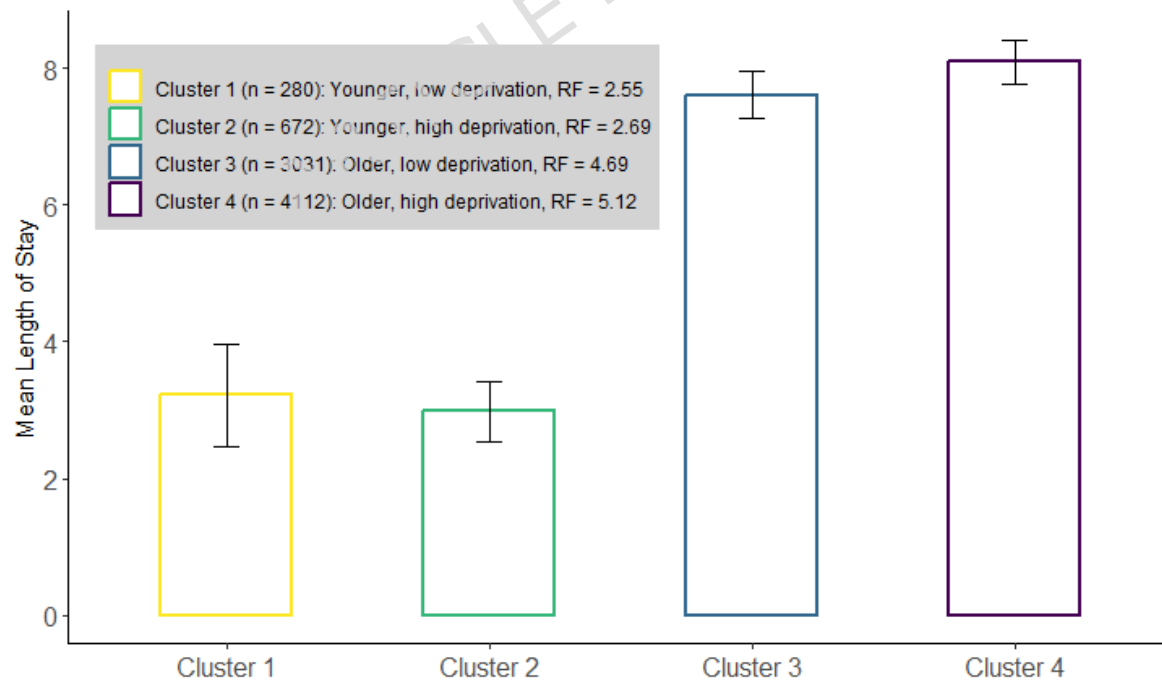


Figure 2. Mean length of stay for each cluster in our analysis with error bars denoting 95% confidence intervals. RF denotes the mean number of risk factors in that cluster.

ANOVA revealed statistically significant group-level differences between clusters ($F=71.17$, $p<.0001$).

Discussion

In this population-based study, obesity, multi-morbidity, and the prescription of Falls Risk Increasing Drugs were the most common falls risk factors. Anemia and vitamin D deficiency were common in the two younger patient groups while cognitive impairment, sensory issues, polypharmacy, a frailty and arthritis were prevalent in the two older patient groups. For both the older and younger cluster pairs, the main distinguishing characteristic was deprivation level. Younger patients with higher levels of deprivation were at higher risk of falls-related hospital admission. In addition to the higher likelihood of presenting to hospital after a fall, the older, multi-morbid patients also had the longest lengths of stay. We observed a relatively high proportion of White patients in the older, high-risk clusters and a relatively high proportion of Asian or British Asian ethnicity in the younger clusters.

While most EHR- based falls research centres on supervised prediction, we use unsupervised, population-level clustering of London-based records to define practical risk segments that can help prioritise reviews. We observed a high prevalence of obesity across all clusters including younger patient groups with a low risk of requiring falls-related care in hospital; these results are consistent with previous research which has shown that obese individuals are more likely to fall but do not have a higher risk of fall-related injuries [29], [30]. Consistent with previous research, we have demonstrated that risk factors for falls tend to aggregate and accumulate with age, leading to different levels of risk for falls requiring hospital-based care [18], [20]. We have shown that this risk is 11-fold in older patients who present with multiple risk factors for falls compared with younger patients with fewer risk factors. Beyond multi-morbidity and frailty, few previous studies have explored whether specific combinations of comorbidities account for a disproportionate use of hospital resources relating to falls [30]. Frailty was indeed highly prevalent in our cohort, but we also observed high proportions of arthritis and diabetes in the clusters at greatest risk of requiring hospital-based care. Although the co-occurrence of these conditions can be explained by older age, focusing risk management efforts on people with this specific comorbidity profile may be an effective approach [18].

Our finding that younger patients with higher levels of deprivation were at higher risk of falls-related hospital admission is meaningful in the context of limited existing evidence evaluating this relationship [19], [31], [32]. This finding is particularly significant when considering the complex interactions between depression, obesity, and fall risk. Depression can impair cognitive and motor functions, while obesity contributes to balance issues and reduced physical activity. The interconnected nature of these factors suggests that targeted interventions addressing mental health, weight management, and socio-economic barriers could potentially reduce fall-related hospitalisation. By understanding these multifaceted risk factors, healthcare providers can develop more comprehensive strategies to prevent falls and improve patient's outcomes, especially among vulnerable younger adults experiencing economic hardship.

We observed that a high proportion of patients in the high-risk clusters were White, whilst Asian or British Asian ethnicity was highly prevalent in the low-risk groups. We observed that a high proportion of patients in the high-risk clusters were White, whilst Asian or British Asian ethnicity was highly prevalent in the low-risk groups. While few studies have considered differences in falls-related hospital resource utilisation between ethnic groups, our results align with pre-existing evidence of differences in fall prevalence between ethnic groups [33]. Potential explanations for these observed differences remain poorly understood. The interaction between ethnicity, deprivation, and clinical risk is complex and likely reflects both differential access to preventive care and variation in the prevalence of underlying conditions across communities. In Cluster 2, which had the highest proportion of Black or Black British patients alongside high deprivation levels, the elevated falls-related hospital risk may partly reflect socio-economic barriers to community-based falls prevention. These interactions warrant further investigation in dedicated analyses with appropriate statistical adjustment.

Each cluster profile suggests a distinct priority for preventive action: older, multimorbid patients (Clusters 3–4) are suited to clinical medication review and multidisciplinary falls assessment, while younger, deprived patients (Cluster 2) are better targeted through community-based and social prescribing approaches. Targeting older patients with specific comorbidity profiles could be more effective for falls prevention than generalized prevention programmes. Integrated Care Boards should concentrate resources towards polypharmacy, particularly among older patients with multi-morbidity, leveraging digital tools to identify and prioritise patients for SMR [34], [35]. For example, a recent study leveraged EHR data to create a prediction tool for fall-related hospital attendance based on multifactorial risk [36]. This kind of prediction tool could be used to prioritise high-risk patients in greatest need of SMR. Our results highlight the need for such a tool to be calibrated to the local population, as we were able to identify homogenous clusters of patients based on risk factor profiles and demographic features specific to NWL population. Population-based approaches using social prescribing interventions directed towards younger patients could target obesity, anaemia, and vitamin D deficiency by addressing lifestyle factors such as healthy eating, physical activity, and spending time outdoors. However, further research is needed to determine the effectiveness of social prescribing in relevant areas such as chronic disease management and frailty prevention [37], [38]. Exercise referral schemes are a promising avenue towards increasing physical activity [39], not only for improving mobility amongst frail populations, but also to reduce prevalence of obesity, itself a falls risk factor [28]. Overall, older, multimorbid segments (Clusters 3–4) can be proactively prioritised for Structured Medication Reviews (with FRID deprescribing prompts) and multidisciplinary falls-clinic referral, while the younger, high-deprivation segment (Cluster 2) is suited to community interventions (social prescribing, weight management, anaemia/vitamin D optimisation). These FRID prompts can be aligned with STOPPFall, the consensus tool for identifying fall-risk-increasing prescriptions in older adults [40]. Although these programmes are widely available across local authorities in England and Wales, further research is required to assess their effectiveness towards health outcomes [41]. Finally, our findings highlight the need for further research to understand how consideration of social deprivation and ethnicity could lead to more effective population level falls risk management strategies in diverse communities such as NWL.

Strengths of this study include the large and diverse sample of adult patients and data from linked patient records across primary and secondary care. The study population is drawn exclusively from North West London, but NWL serves a diverse population of 2.1 million people annually, with sociodemographic characteristics summarised in Table S3. Moreover, the methodology, rather than the specific cluster parameters, is intended to be transferable to other Integrated Care Systems or health systems with comparable linked EHR infrastructure. We sought to add to the evidence base by evaluating patterns of risk factors and how they interact to increase the risk of a falls-related hospital encounter; it was not possible to draw conclusions about the effects of individual risk factors with our study design. The prevalence of certain falls risk factors could be underestimated in our dataset, which only contains records of falls risk factors where clinically coded. It is likely that some risk factors are coded inconsistently – for example, we observed a low rate of coded mobility problems despite a high prevalence of frailty. We attempted to mitigate against variation in coding practices by using standard code lists from the QOF and the eFI, where possible. Due to coding practices, our analyses may not have captured all patients attending the ED after a fall; we suspect that falls are not coded consistently in the ED as many of the patients coded for a fall during their initial hospital admission were not coded for a fall while in the ED. Healthcare cost data were not available within the current data access approval; future studies with access to HRG-level cost data could provide a more complete picture of the economic burden by cluster. Although more granular information (e.g., fracture types and other injuries) were available, these were intentionally aggregated into clinically defined falls-related risk factors (Table S1) to ensure consistency with established clinical assessment frameworks and to reduce sparsity across low-frequency subcategories. Given that falls appear to be undercoded in the ED, our outcome measure likely underestimates the true burden of falls-related hospital use. A formal sensitivity analysis varying the outcome definition was not feasible within the current

study design and will be explored for future work. Since only GP prescriptions are included in the WSIC database, our analyses do not include medications prescribed in other settings. Finally, we used the eFI to determine frailty prevalence as the tool uses validated code lists. However, the eFI is known to overestimate frailty [42]. As with all EHR-based studies, the findings are subject to limitations arising from coding completeness, consistency, and accuracy. Formal stability analysis of the cluster solution (e.g., via bootstrap resampling) was not conducted and represents a direction for future validation work. These findings should be interpreted in the context of the study design. Prospective evaluation of targeted interventions based on these risk profiles will be needed to establish their effectiveness in reducing falls-related hospital use. Formal stability analysis of the cluster solution (e.g., via bootstrap resampling) was not conducted and represents a direction for future validation work. **Future work also should seek to replicate this methodology in rural populations and health systems with different demographic profiles. Integration of real-time monitoring data such as sensor-based measurements alongside EHR-derived risk profiles represents a promising direction for improving the precision of individual-level risk assessment.**

Conclusions

Population-level falls risk factor profiles can be determined using coded data from linked EHR. This study has identified distinct clusters of adult patients in the NWL population based on sociodemographic characteristics and falls risk factors that are coded in patients' EHR. Older, multi-morbid patient groups with a high prevalence of arthritis and diabetes are much more likely to require hospital-based care and stay in hospital longer following a fall relative to younger, fitter population clusters. Amongst younger patient groups, patients with higher deprivation are more likely to require hospital-based care than patients with lower deprivation. This study highlights the potential of unsupervised machine learning on routinely collected EHR data as a scalable, practical approach to uncovering high-risk profiles and informing more effective, data-driven falls prevention strategies.

References

1. Montero-Odasso, M. et al. World guidelines for falls prevention and management for older adults: a global initiative. *Age Ageing* 51, afac205 (2022). <https://doi.org/10.1093/ageing/afac205>
2. Gillespie, L. D. et al. Interventions for preventing falls in older people living in the community. *Cochrane Database Syst. Rev.* CD007146 (2012). <https://doi.org/10.1002/14651858.CD007146.pub3>
3. Tang, C. T. L. et al. Secular trends in fall-related hospitalizations in adolescents, youth and adults: a population-based study. *Lancet Reg. Health West. Pac.* 12, 100183 (2021). <https://doi.org/10.1016/j.lanwpc.2021.100183>
4. Montero-Odasso, M. M. et al. Evaluation of clinical practice guidelines on fall prevention and management for older adults: a systematic review. *JAMA Netw. Open* 4, e2138911 (2021).
5. National Institute for Health and Care Excellence. Assessment and prevention of falls in older people (NICE, 2013).
6. Haagsma, J. A. et al. Falls in older aged adults in 22 European countries: incidence, mortality and burden of disease from 1990 to 2017. *Inj. Prev.* 26(Suppl 2), i67-i74 (2020).
7. Morris, R. & O'Riordan, S. Prevention of falls in hospital. *Clin. Med.* 17, 360-362 (2017). <https://doi.org/10.7861/clinmedicine.17-4-360>
8. Heikkilä, A. R. N. et al. Consequences of inpatient falls in acute care: a retrospective register study. *J. Patient Saf.* 20, 340-344 (2024).

<https://doi.org/10.1097/PTS.0000000000001230>

9. Office for National Statistics. Overview of the UK population (2021).
10. Daniels, H. et al. The value of routinely collected data in evaluating home assessment and modification interventions to prevent falls in older people: systematic literature review. *JMIR Aging* 4, e24899 (2021).
11. National Institute for Health and Care Excellence. Falls in older people: quality standard (NICE, 2015).
12. Jehu, D. A. et al. Risk factors for recurrent falls in older adults: a systematic review with meta-analysis. *Maturitas* 143, 23–28 (2021).
13. Li, Y. Risk factors for falls among community-dwelling older adults: a systematic review and meta-analysis. *Front. Med.* 7, 222 (2020).
14. Sherrington, C. et al. Exercise for preventing falls in older people living in the community. *Cochrane Database Syst. Rev.* CD013789 (2019). <https://doi.org/10.1002/14651858.CD013789.pub2>
15. Lewis, S. R. et al. Population-based interventions for preventing falls and fall-related injuries in older people. *Cochrane Database Syst. Rev.* CD013789 (2024).
16. Goh, D. F. et al. Population health management. In: *Clinical Decision Support Systems* (Elsevier, 2023).
17. Archer, L. et al. Development and external validation of a risk prediction model for falls in patients with an indication for antihypertensive treatment. *BMJ* 379, e070918 (2022). <https://doi.org/10.1136/bmj-2022-070918>
18. Finch, C. F. et al. Identifying clusters of falls-related hospital admissions to inform population targets for prioritising falls prevention programmes. *Inj. Prev.* 21, 254–259 (2015).
19. Court-Brown, C. M. et al. The relationship of fall-related fractures to social deprivation. *Osteoporos. Int.* 22, 1211–1218 (2011). <https://doi.org/10.1007/s00198-010-1315-1>
20. Ek, S. et al. Risk profiles for injurious falls in people over 60: a population-based cohort study. *J. Gerontol. A Biol. Sci. Med. Sci.* 73, 233–239 (2018).
21. Bottle, A. et al. How an electronic health record became a real-world research resource. *BMC Med. Inform. Decis. Mak.* 20, 71 (2020). <https://doi.org/10.1186/s12911-020-1082-7>
22. NHS England Digital. Quality and Outcomes Framework (QOF) business rules v47.0 2022–2023 (2023).
23. Healthcare Improvement Scotland. Read codes used by the electronic frailty index in SPIRE (2019).
24. Health Data Research UK. Phenotype Library (accessed 2025).
25. Masnoon, N. et al. What is polypharmacy? A systematic review of definitions. *BMC Geriatr.* 17, 230 (2017).
26. National Falls Prevention Coordination Group. National falls prevention strategy (2023).
27. Lever, J., Krzywinski, M. & Altman, N. Principal component analysis. *Nat.*

Methods 14, 641–642 (2017).

28. MacQueen, J. Some methods for classification and analysis of multivariate observations. *Proc. 5th Berkeley Symp. Math. Statist. Probab.* 281–297 (1967).
29. Mitchell, D. et al. Quality improvement in vascular surgery: the role of comparative audit and Vascunet. *Eur. J. Vasc. Endovasc. Surg.* 49, 1–3 (2015).
30. Gnudi, S. et al. Does obesity increase the risk and severity of falls in people aged 60 years and older? *J. Gerontol. A Biol. Sci. Med. Sci.* 75, 952–960 (2020).
31. West, J. et al. Do rates of hospital admission for falls and hip fracture in elderly people vary by socio-economic status? *Public Health* 118, 576–581 (2004).
32. Johnson, N. A. & Dias, J. The effect of social deprivation on fragility fracture of the distal radius. *Injury* 50, 1232–1236 (2019).
33. Wehner-Hewson, N. et al. Racial and ethnic differences in falls among older adults: a systematic review and meta-analysis. *J. Racial Ethn. Health Disparities* 9, 2427–2440 (2022).
34. Li, L. et al. Structured medication reviews for patients with polypharmacy in primary care. *J. Comorbidity* 15, 20542704251325056 (2025).
35. Woodcock, T. et al. Polypharmacy in primary care: a population-based retrospective cohort study. *PLOS ONE* 19, e0308624 (2024).
36. Walker, L. E. et al. A plan to reduce overprescribing to make patient care better and safer. *J. Multimorb. Comorbidity* 12, 263355652211454 (2022).
37. Walker, L. E. et al. The DynAIRx project protocol: artificial intelligence for dynamic prescribing optimisation. *J. Multimorb. Comorbidity* 12, 26335565221145493 (2022).
38. Harrison, J. K. et al. Development and external validation of the eFalls tool. *Age Ageing* 53, afae057 (2024).
39. Htun, H. L. et al. Effectiveness of social prescribing for chronic disease prevention. *J. Epidemiol. Community Health* 77, 265–276 (2023).
40. Seppälä, L. J. et al. STOPPFall: a Delphi study on fall-risk-increasing drugs. *Age Ageing* 50, 1189–1199 (2021).
41. Smith, T. O. et al. Social prescribing programmes to prevent or delay frailty in community-dwelling older adults. *Geriatrics* 4, 65 (2019).
42. A. Broad, B. Carter, S. McKelvie, and J. Hewitt, “The Convergent Validity of the Electronic Frailty Index (eFI) with the Clinical Frailty Scale (CFS)”, *Geriatrics*, vol. 9;5(4):88, Nov, 2020.

Author contributions statement

R.L., R.T., and C.C. designed the study, curated the WSIC dataset and code lists, implemented the PCA+k-means pipeline, performed the statistical analyses, created the figures and tables, and drafted the manuscript. E.Y. and Z.L. refined the outcome definitions, reviewed and validated the machine-learning methodology, led the literature review and related-work synthesis, and reviewed results for face validity. Y.A. obtained approvals, coordinated data access and information governance, and supported cohort construction. B.G. advised on methodology, robustness checks, and sensitivity analyses. C.M. provided expertise on falls coding, FRIDs, and critically revised the manuscript. E.M. provided overall supervision, contributed to study design and interpretation, and critically revised the manuscript. All

authors reviewed and approved the final version.

Data access and approvals

Informed consent was not required for this research as it involved the secondary use of anonymised, routinely collected electronic health record data accessed within a secure research environment. This research database is approved by the Health Research Authority with favourable ethics approval given by the West Midlands Solihull Research Ethics Committee (reference 18/WM/0323; IRAS project ID 252449). Access to the research database, to undertake this study, was provided by the NW London Sub-Data Research Access Group (SDRAG Ref ID 342). All data used in this work were fully anonymised before analysis.

Data and code availability

The datasets generated and/or analysed during the current study are not publicly available due to institutional, regulatory and data governance restrictions. Researchers may contact the iCARE Secure Data Environment and the corresponding author for further information regarding access procedures and required institutional and regulatory approvals.

Conflict of interest

The authors declare no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

ARTICLE IN PRESS