communications medicine

A Nature Portfolio journal

Article

Article in Press

https://doi.org/10.1038/s43856-026-01917-8

Challenges of clustering disease trajectories in people with multiple long-term conditions using data from 7.2 million patients

Check for updates

A list of authors and their affiliations appears at the end of the paper

Abstract

Background Generating clusters of people with similar patterns of Multiple Long-Term Conditions (MLTC) could help target healthcare services for specific patient groups. We aimed to generate data-driven clusters of patients based on their disease trajectories and assess clinical interpretability and associations with future health outcomes.

Methods We used structured general practice data from the Clinical Practice Research Datalink Aurum, linked to Hospital Episode Statistics data on emergency department (ED) attendances and hospital admissions, including all adults registered on 1st January 2015. For each patient, we generated a vector embedding representing their trajectory of diseases developed over time, using a transformer model, followed by clustering using k-means. Patients were subsequently followed up for 1 year to estimate associations of cluster membership with ED attendance, hospitalisation and mortality, compared to patients with no long-term conditions. We also evaluated the use of the clusters for predicting these outcomes compared with using number of long-term conditions, individual diseases or embeddings alone.

Results Analysis included 5,981,091 (82.1%) patients with MLTC and 1,304,119 (17.9%) with no chronic conditions. We identified eight clusters of patients with MLTC as optimal. The largest, representing 21.3% of the population included strong contributions from cardio-kidney-metabolic conditions, but there was substantial overlap in prevalent conditions across clusters. Although large differences were found between clusters in 1-year odds of ED attendance, hospitalisation and mortality, with the highest odds in cardio-kidney-metabolic clusters, clusters explained only between 1.2-3.1% of total variance, and associations were substantially attenuated after adjustment for age, sex, ethnicity and deprivation. Clusters performed substantially worse at predicting outcomes than using the individual diseases or embeddings, and performed worse than a person's count of long-term conditions at predicting ED attendance and hospitalisation.

Conclusions We generated eight data-driven clusters of adults with similar trajectories of MLTCs. Interpretation of the clusters was challenging due to overlap of the dominant conditions across clusters, while clusters explained little of the variation in 1-year health outcomes. Our findings highlight challenges in translating data-driven clusters of disease trajectories into interpretable groups that can inform future care needs.

e-mail: thomas.beaney@imperial.ac.uk

Multimorbidity, also known as Multiple Long-Term Conditions (MLTC), is a health state defined as the co-existence of two or more Long-Term Conditions (LTCs) in one person[1]. With population ageing and the increasingly effective management of many LTCs, the prevalence of MLTC is a growing problem for health systems globally[2]. Having MLTC is associated with increased mortality[3], greater use of health services[4], poorer quality of life[5], and challenges in navigating healthcare services[6]. One of the central challenges in developing strategies to care for people with MLTC is its heterogeneity, with tens of thousands of unique combinations of co-existing conditions[7]. Over the last decade, unsupervised clustering methods have been widely used as a data-driven approach to distinguish common patterns of MLTC, consistently identifying cardiometabolic and mental health clusters[8,9].

While clustering of diseases has been the predominant approach in the literature[8], disease clusters may not map clearly to people, who will often have diseases in many clusters[10]. A growing body of research instead aims to create clusters of people who share similar diseases[11,12]. Applied to healthcare, these approaches are analogous to "segmentation" methods, an approach "to categorising individuals with similar characteristics with the aim of identifying common care needs and designing services tailored to those in each segment"[13]. For clusters to inform the design of care pathways, they should be clinically interpretable, while to inform risk stratification, they should capture meaningful variation in patient outcomes.

Although many studies have applied segmentation in healthcare, most create broad groups, such as a cluster of people with any chronic illness, which does not capture the diversity within MLTC[14–16]. Furthermore, most clustering studies represent MLTC at a single point in time[8,9]. This ignores the sequence in which conditions develop, which may provide valuable information on underlying patterns of disease progression[17]. Sequence-based approaches, such as transformer models, have demonstrated strong performance in capturing disease trajectories over time and predicting health outcomes[17,18], but whether grouping these representations into clusters produces clinically meaningful and informative patient groups remains uncertain.

Here, we aim to generate data-driven clusters of patient disease trajectories and assess their interpretability and ability to capture variation in future health outcomes. Using a pipeline incorporating a transformer model and k-means clustering, we first generate unsupervised clusters of people with MLTC, based on their chronological patterns of disease development. We then characterise each cluster by their prevalent LTCs and sociodemographic profiles. Finally, we investigate the extent to which clusters capture variation in future healthcare utilisation and mortality over one-year follow-up.

Methods

Data sources and data cleaning

This cohort study uses the Clinical Practice Research Datalink (CPRD) Aurum dataset, which collates patient-level data from general practice electronic health records (EHRs) in England, and is nationally representative[19]. Secondary care data on emergency department (ED) attendances and hospital admissions were sourced from Hospital Episode Statistics (HES), and death registration data from the Office for National Statistics. Linkage was performed by NHS Digital using a pseudonymised patient identifier. Only patients eligible for linkage were included, resulting in 1.2% being excluded. Our data extract includes only patients marked as 'research acceptable' by CPRD[20]. We included patients aged ≥18 years who were registered on 1st January 2015. We required that all patients were registered for at least one year before 1st January 2015 to allow sufficient time for data input[21], and that all patients had at least one year of follow-up. Two cohorts were constructed: the first included those with two or more LTCs (see definition below), for clustering, and the second included those with no LTCs to provide a consistent comparator group.

As a marker of socioeconomic deprivation, we used the Index of Multiple Deprivation 2019, linked to each patient based on their geographic area of residence[22]. This is a composite area-level measure capturing a range of socioeconomic factors. Ethnicity was categorised using standard NHS categories (see Supplementary Material p.2). Deaths can be recorded in both CPRD and in ONS mortality statistics – in cases where a death was recorded only in one source, or was recorded in both, but with different dates, we reconciled the sources adapting an algorithm developed by Delmestri and Prieto-Alhambra (see Supplementary Material p.2)[23]. In all analyses, missing sociodemographic data were encoded as missing indicator variables.

Disease definitions

We used code lists derived in earlier research representing 212 LTCs relevant to MLTC (Supplementary Material p.3-9)[24–26]. These code lists map 9,462 diagnostic codes (the 'Medcode' clinical ontology in CPRD) entered during clinical encounters to the 212 LTCs. For each patient, a sequence of diseases was constructed according to the temporal order of clinical codes in their EHR, allowing for repetition. For example, a disease sequence for a patient might read "hypertension, rheumatoid arthritis, hypertension, chronic kidney disease, gout, gout". For each patient, their entire history of LTCs was included, from birth up until 31st December 2014. In CPRD, each code has an 'observed' date and an 'entry' date; we used the former to define the order of diseases, as diagnostic codes may be back-dated to the date of observation.

Generating a patient representation

Generating clusters requires a quantifiable representation of the disease sequences as an input. To produce these representations, we generated an unsupervised 'vector embedding' for each patient. Here, we generated the representation using a transformer architecture, EHR-BERT, which we developed in earlier work[17]. In brief, this algorithm, adapted from large language models used for text inputs[27], uses the sequence of each patient's diseases as an input to try to predict randomly 'masked' diseases within the sequence. In doing so, it learns a vector representation of each person's disease sequence. The information on disease sequences is supplemented with information on age, gender, ethnicity, socioeconomic deprivation and calendar year for each code. While gender, ethnicity and deprivation are fixed throughout, age and calendar year associated with each code will vary over time, allowing the model to learn the impact of these factors on disease sequences. More information is provided in Beaney et al. (2024), but the result is the generation of a vector of length 288 for each patient, representing their unique disease trajectory[17]. Our previous work indicated that this representation captures deep contextual information that offers better predictive power than methods of representation using only the co-occurrence of diseases across a range of clinical outcomes[17].

To assess the impact of sociodemographic embeddings within the transformer architecture, we conducted a sensitivity analysis applying the same clustering pipeline to embeddings generated from a simpler Med-BERT architecture derived from our earlier work[17,28]. This model generates embeddings using the sequence of disease codes only, without sociodemographic information.

Clustering

We use the vector embeddings as input to k-means clustering for the cohort with MLTC. We initially experimented with different clustering methods, such as density-based and graph-based clustering[29], but it was not computationally feasible to run these on the entire dataset. For practical use, we hypothesised that between four and 40 clusters would balance meaningful separation of patient phenotypes with interpretability, and ran k-means clustering for all values of k clusters between four and 40. This was implemented in the scikit-learn Python package, using the Lloyd algorithm and iterating 1000 times with different random centroids[30]. We selected the optimal value of k clusters, using both Davies-Bouldin and Calinski-Harabasz scores[31]. With high-dimensional data, separation of clusters can be challenging using methods based on Euclidean distances such as k-means[32]. As a sensitivity analysis, we tested two variants of the 288-length input embedding with the number of dimensions reduced via Principal Component Analysis (PCA): one explaining 95% of the variance (96 dimensions)

Table 1 | Characteristics of the study population for those with MLTC and those with no LTCs

Patient characteristic	With MLTC		With no LTCs	
	Total	Percent	Total	Percent
Age (years)				
Mean (SD)	54.2 (18.1)		36.0 (13.1)	
18–29	635,951	10.6%	483,464	37.1%
30–39	738,864	12.4%	380,631	29.2%
40–49	1,059,467	17.7%	239,663	18.4%
50–59	1,183,828	19.8%	127,768	9.8%
60–69	1,048,679	17.5%	48,889	3.7%
70–79	777,891	13.0%	15,992	1.2%
80+	536,411	9.0%	7712	0.6%
Gender				
Female	3,172,062	53.0%	509,105	39.0%
Indeterminate	54	<0.1%	15	<0.1%
Male	2,808,975	47.0%	794,999	61.0%
Ethnicity				
White	5,167,461	86.4%	850,120	65.2%
South Asian	351,992	5.9%	127,547	9.8%
Black	207,090	3.5%	67,234	5.2%
Other	77,371	1.3%	72,885	5.6%
Mixed	59,219	1.0%	24,079	1.8%
Missing	117,958	2.0%	162,254	12.4%
IMD decile				
1 (least deprived)	681,578	11.4%	119,686	9.2%
2	623,792	10.4%	113,088	8.7%
3	634,813	10.6%	129,157	9.9%
4	631,560	10.6%	123,984	9.5%
5	564,937	9.4%	127,624	9.8%
6	592,094	9.9%	138,709	10.6%
7	589,598	9.9%	147,314	11.3%
8	565,316	9.5%	151,630	11.6%
9	566,781	9.5%	137,429	10.5%
10 (most deprived)	525,831	8.8%	114,218	8.8%
Missing	4791	0.1%	1280	0.1%
Total	**5,981,091**	**82.1%**	**1,304,119**	**17.9%**

and the second explaining 90% of the variance (71 dimensions) of the original embedding.

Descriptive analysis of clusters

To account for conditions which are more prevalent in the population, we calculated the ratio of observed to expected number of people with an LTC in each cluster. Here, the expected number, E is given by the formula $E = N * P$, where N is the number of people in the cluster, and P is the prevalence of the disease in the whole population. We then calculated a Z-score for the difference between observed and expected prevalence of each LTC in each cluster, and the two-tailed p-value under the null hypothesis of the difference being zero. If the observed-to-expected ratio was greater than one and the p-value $p < 0.00024$ (applying a Bonferroni correction to account for multiple testing), we retained the disease in the cluster. We visualised the diseases in each cluster using word clouds, where larger text size indicates relative prevalence of a condition in the cluster, using the Python wordcloud package (version 1.9.2)[33]. Thus, the inclusion of a disease in a cluster represents it occurring more than expected by chance given the population prevalence, and the text size represents prevalence relative to other conditions within the same cluster.

Associations with healthcare utilisation and mortality

For each cluster (defined as of 31st December 2014), we calculated the percentage of patients i) presenting to ED at least once, ii) with one or more emergency hospital admissions, and iii) who died during one year of follow-up through 2015 (i.e., in the year after information input into the clustering was captured). We selected a one-year follow-up period as a practical planning horizon commonly used within health services to inform service delivery. For each outcome, we calculated the total variance both within and between clusters. To examine the strength of association of each cluster with each outcome, we used logistic regression, incorporating the cluster assignment of a patient as a categorical variable using Stata version 17[34]. The reference group included patients with no LTCs as a fixed baseline that did not require selection of a reference cluster; a sensitivity analysis compared outcomes only within the MLTC population using the youngest cluster as the reference group. To assess the degree to which cluster associations were explained by sociodemographics and number of LTCs, we constructed three models:

1. Model 1: unadjusted
2. Model 2: adjusted for age, gender, ethnicity and deprivation decile
3. Model 3: adjusted as for model 2, plus the number of LTCs per person

Age category, gender, ethnicity and deprivation deciles were incorporated as categorical variables. In Model 3, the number of LTCs was incorporated as a continuous variable, along with a quadratic term, which was found to be statistically significant across all models.

Predicting healthcare utilisation and mortality

We assessed the performance of the clusters to predict the same three outcomes over 1-year, training a logistic classifier with L2 penalisation and four-fold cross-validation in scikit-learn[30]. Due to the small number of deaths, for mortality we included only adults aged ≥ 60 years. Clusters were input as dummy variables. Models were evaluated using the Area Under the Curve (AUC). We compared performance against three models with alternative inputs:

1. Number of LTCs (as a continuous variable)
2. Presence or absence of each disease (as binary indicators)
3. The original patient embeddings only (as a vector of continuous values)

For each input, we compared performance using: i) the input alone; ii) inclusion of sociodemographics (age, gender, ethnicity and deprivation decile); iii) additional inclusion of the count of LTCs per person.

Patient and public involvement

This work was supported by a Patient and Public Advisory Group (PPAG) comprising three people with lived experience as a patient or carer for a person with MLTC who gave informed consent to participate. Three workshops were conducted over the course of the work, informing the selection of the 212 LTCs and outcomes, and the interpretation of results. Key findings from our PPAG are reported in the Discussion.

Ethical approval

Data access to CPRD and ethical approval was granted by CPRD's Research Data Governance Process on 28th April 2022 (Protocol reference: 22_001818) and with linkage to HES data on 6th March 2023 (Protocol reference: 22_002481).

Results

Study population

A total of 5,981,091 (82.1%) patients with MLTC and 1,304,119 (17.9%) with no LTCs were included. Those with MLTC were older, with a mean (SD) age of 54.2 years compared with 36.0 years in those with no LTCs (Table 1). Gender profiles also differed, with more females (53.0%) in those

with MLTC and more males (61.0%) in those with no LTCs. People with no LTCs were more likely to have missing ethnicity (12.4% versus 2.0%), suggesting that those with more LTCs may have more opportunities for documenting ethnicity in EHR data. Among people with MLTC, there were slightly more patients in the least deprived deciles, in contrast to those with no LTCs, where there were slightly more patients in the most deprived deciles.

Clusters of people with MLTC

We generated clusters of people with MLTC using k-means. Based on both the Calinski-Harabasz and Davies-Bouldin scores, local minima were identified at resolutions of five, eight, fourteen, seventeen, 27 and 37 clusters (Supplementary Fig. 1). These six optima were investigated further. A Sankey plot of the assignment of people between the six partitions demonstrated a loosely hierarchical pattern, with most conditions remaining with other conditions in the same cluster as the number of clusters reduced (Supplementary Fig. 2).

Two sensitivity analyses using embeddings following dimensionality reduction with PCA (explaining 90 or 95%) gave different optimal sets of clusters (both at six, ten, 24 and 34), but the contents of the clusters were similar across partitions, suggesting minimal impact on the composition of the clusters (Supplementary Figs. 3, 4).

Disease patterns across clusters

Of the six optimal partitions, we focus on the eight-cluster partition. Eight clusters provided more granular insights than only five clusters, but with higher numbers of clusters of fourteen, 27 or 37, it became less clear which diseases made a cluster unique. For example, in the fourteen-cluster partition, clusters 1, 2, 4, and 6 shared similar cardio-kidney-metabolic conditions, while clusters 10, 12, and 14 shared high prevalences of acne, allergic and chronic rhinitis, asthma and dermatitis. The most common diseases in each cluster, which were statistically more likely to occur than expected, are shown in Fig. 1 (see Supplementary Figs. 5–9 for other optima).

Cluster 1 (21.3% of the population) and cluster 5 (9.8%) had strong contributions from cardio-kidney-metabolic conditions, including diabetes, hypertension and raised cholesterol measurements. Cluster 2 (15.2%) was made up of mostly allergic, skin and pain conditions. Musculoskeletal conditions contributed significantly to cluster 3 (16.0%), along with other conditions affecting different body systems. Cluster 4 (15.1%) was made up of four measurements of raised cholesterol and triglyceride levels. Cluster 6 (8.2%) included mostly mental health conditions, with some pain disorders, and Cluster 7 (8.1%) had high contributions from obesity, asthma, and low high-density lipoprotein (HDL)- cholesterol. Finally, cluster 8 (6.3%) included high contributions from asthma, allergic and chronic rhinitis and dermatitis. There was strong overlap between clusters, with raised cholesterol measurements dominant across Clusters 1, 4, 5, and 7, while asthma appeared strongly across Clusters 7 and 8.

Sociodemographic characteristics of people in each cluster

There were large differences in the age distributions among people in each cluster, with the youngest median age in Cluster 2 (31 years) and oldest in Cluster 5 (70 years) (Table 2). Gender distributions also varied by cluster, with Cluster 7 including the highest percentage of females (74.4%). The relative age distributions by gender were similar within each cluster (Supplementary Fig. 10), but males in Cluster 4 (lipid-related disorders) tended to be younger than females in Cluster 4. The distribution of ethnicity also varied across clusters, with the lowest percentage of people of White ethnicity in Cluster 2 (78.3%), made up of allergic, skin and pain conditions, and the highest in Cluster 6 (92.4%), with a prominence of mental health conditions (Table 2). Similarly, there were differences across deprivation deciles, with the lowest percentage of people living in the most deprived areas in Cluster 3 (6.2%) and highest in Clusters 6 and 7 (12.1% and 12.0% respectively).

Associations with healthcare utilisation and mortality

Clusters explained only a small percentage of overall variance in outcomes: 1.2% for any ED attendance, 3.1% for any emergency admission and 1.8% for mortality over 1 year. Using the larger set of 37 clusters, corresponding percentages remained low (1.4%, 3.7% and 2.5% for each outcome, respectively) (Supplementary Table 1). People in cluster 1 had the highest odds of ED attendance, emergency admission and mortality over 1 year (Figs. 2–4) in unadjusted models (model 1). However, after adjustment for sociodemographics and number of LTCs (model 3), cluster 5 had the highest odds of attending ED. For all clusters, effect sizes were significantly attenuated following adjustment, with a small range of values between outcomes.

Predicting outcomes

We assessed cluster performance for predicting the same three outcomes. Using the 8-clusters alone, predictive performance was substantially worse than using either individual diseases or the embeddings from which they were derived, although differences were attenuated once sociodemographic information was also included as an input (Table 3). The clusters also performed worse than using a person's LTC count alone for predicting ED attendance and hospitalisation, but slightly better for mortality. In a sensitivity analysis examining cluster performance for predicting mortality across the range of cluster sizes from four to 40, there was only a modest increase in the AUC (from 0.781 to 0.799; Supplementary Fig. 16).

Sensitivity analysis using clusters generated without sociodemographic information

We generated embeddings using a simpler Med-BERT architecture which learns from disease sequences without linked sociodemographics, identifying different optimal partitions (Supplementary Figs. 17–20). Comparing 8-clusters for consistency with the primary analysis, both approaches generated two cardio-kidney-metabolic clusters and a lipid cluster. However, the cluster of musculoskeletal conditions did not emerge using Med-BERT, and instead two mental health clusters were formed (Supplementary Fig. 19). In general, clusters derived from Med-BERT embeddings explained slightly less variation in healthcare utilisation than those from EHR-BERT but similar or slightly more variation in mortality (Supplementary Table 3). For prediction, the raw embeddings and clusters derived from Med-BERT performed consistently worse than those from EHR-BERT, but once sociodemographic factors and number of LTCs were included as inputs in the prediction model, performance was similar (Supplementary Table 4).

Discussion

We generated eight data-driven clusters of adults with MLTC, based on 6 million patients' sociodemographic characteristics and disease trajectories recorded in GP records. There were large differences between clusters in sociodemographic profiles and prevalent diseases. Clusters 1 and 5 included older adults with cardio-kidney-metabolic conditions, reflecting shared causes and pathophysiology between these diseases[35], while Cluster 2 included younger adults with skin and allergic conditions. Although people in Clusters 1 and 5 used healthcare services most frequently and had the highest risk of mortality over one year, much of this association was explained by their older average age and higher number of LTCs, rather than the specific mix of conditions, with clusters explaining little of the overall variability in outcomes. Furthermore, clusters performed substantially worse at predicting 1-year outcomes compared with information on either the individual diseases or embeddings from which they were derived, demonstrating the potential loss of information for risk stratification when aggregating individual diseases into clusters.

Ultimately, clustering approaches are useful if they support the design of care pathways or allow identification and subsequent intervention for patients at high risk. For example, well-defined clusters might allow for the creation of specific services tailored to each cluster that would not be feasible to create for every individual, while supporting a shift away from single-

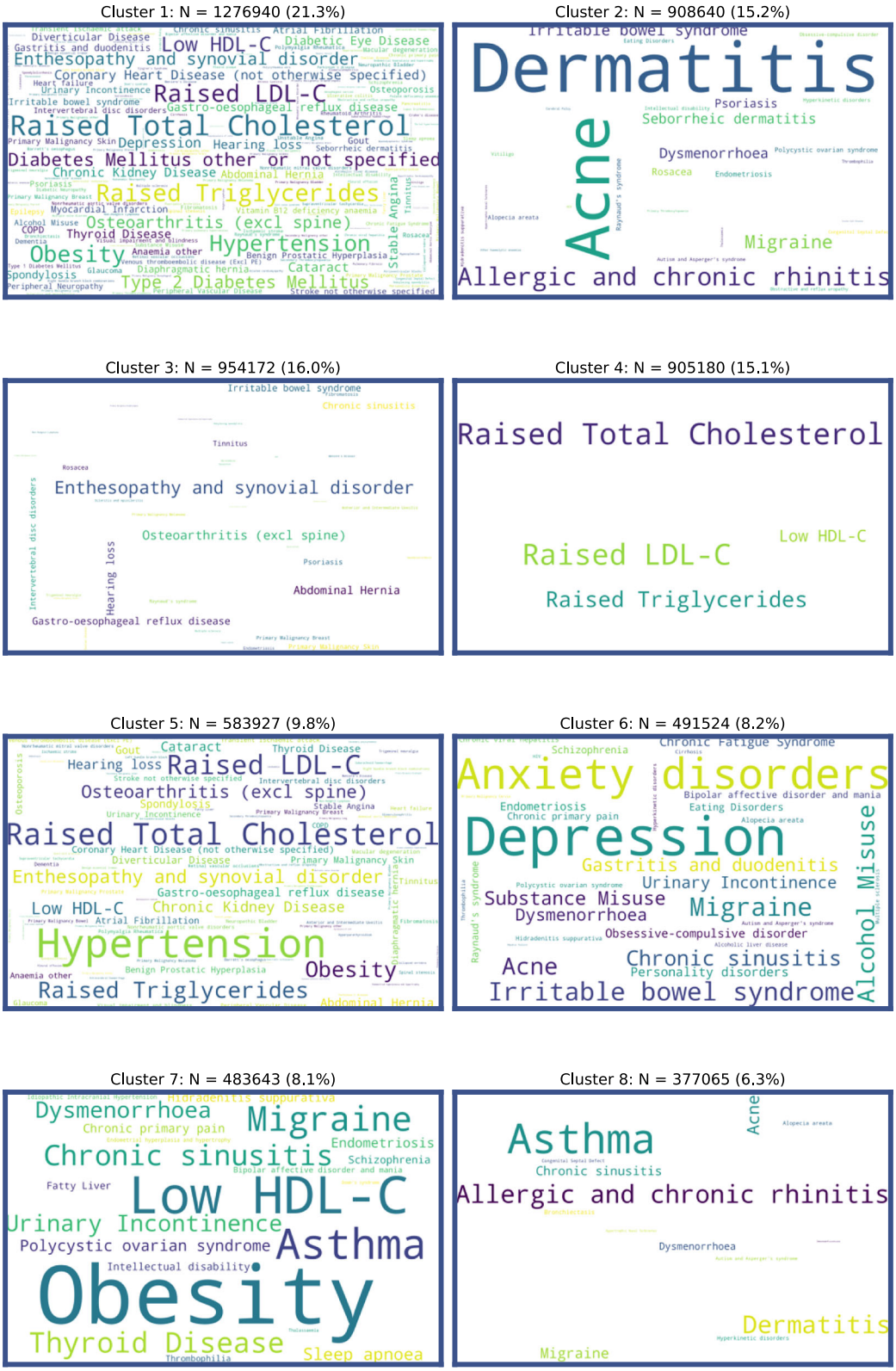


Fig. 1 | Disease clouds of most common diseases in each cluster for the 8-cluster partition. Diseases included only if prevalence is significantly higher than expected by chance. Text size indicates the relative prevalence of a disease within the cluster. For prevalence of each disease in each cluster, see Supplementary Data 1.

disease care. Alternatively, if a cluster enables accurate risk stratification, this might allow clinicians to initiate a review or target treatments. From our analysis, two prominent challenges emerged that impact the implementation of findings in clinical practice: 1) interpretability; 2) discrimination of future outcomes.

Identifying the features distinguishing each cluster was challenging. While some of our identified clusters map onto known relationships, such as cardio-kidney-metabolic conditions in clusters 1 and 5[35], connections within other clusters were not well-established. Overlap in the conditions across clusters also introduces complexity to interpretation, which increased

Table 2 | Characteristics of people in each cluster

Cluster	Median (IQR)age (years)	Gender(% female)	Ethnicit(% White)*	Deprivaton(% in mostdeprived decile) Ψ	Median (IQR)numbr of LTCs
Cluster 1	69 (58–79)	50.0	88.8	10.4	10 (7–13)
Cluster 2	31 (25–39)	62.1	78.3	8.9	3 (2–4)
Cluster 3	56 (49–66)	45.1	90.7	6.2	4 (2–5)
Cluster 4	58 (50–67)	41.8	83.1	6.6	5 (3–7)
Cluster 5	70 (61–79)	52.4	89.2	7.1	7 (5–10)
Cluster 6	45 (36–54)	64.3	92.4	12.1	5 (3–7)
Cluster 7	47 (38–56)	74.4	83.3	12.0	5 (3–7)
Cluster 8	38 (27–50)	47.5	86.5	9.1	4 (3–6)

*: see Supplementary Fig. 11 for a breakdown of clusters across all ethnic groups.
Ψ: see Supplementary Fig. 12 for a breakdown of clusters across all deprivation deciles.

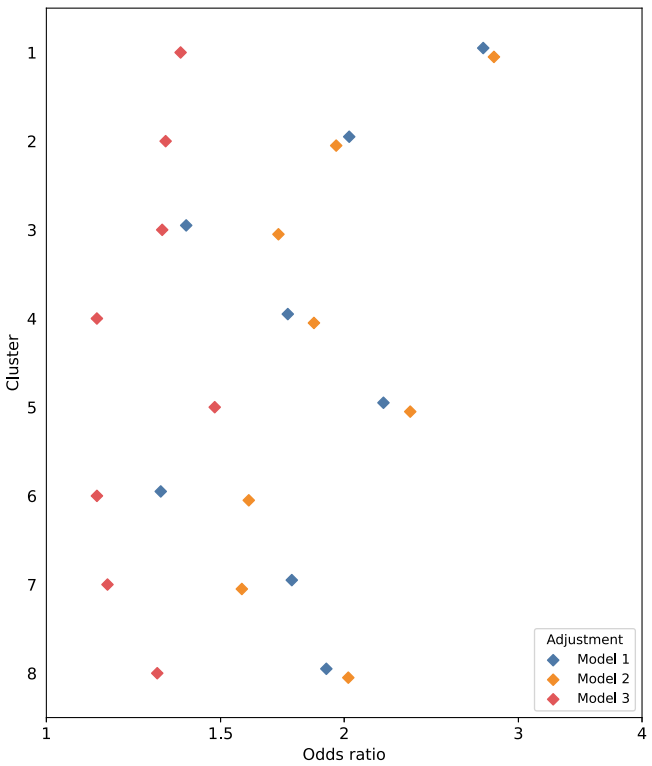


Fig. 2 | Odds ratios for any ED attendance over one year for each cluster, compared to people with no LTCs. Model 1: unadjusted; Model 2: adjusted for sociodemographics; Model 3: adjusted for sociodemographics and number of LTCs per person. Vertical line at 1 represents OR for people with no LTCs. Point estimates and confidence intervals shown in Supplementary Table 2. A comparison of outcomes among those with MLTC only is shown in Supplementary Fig. 13.

as the number of clusters increased. Methodological choices, including the inclusion of 212 LTCs and transformer-based architectures, are likely to contribute to this complexity. However, use of the simpler Med-BERT architecture produced clusters which were similarly challenging to interpret. Furthermore, other studies using fewer diseases and simpler cross-sectional methods have also found strong overlap across clusters. For example, Robertson et al. found that hypertension was among the three most prevalent conditions in seven out of ten clusters of people based on co-occurrence of 30 diseases[11]. Thus, the challenge of interpreting unsupervised clusters of patients may reflect their inherent heterogeneity within the population, rather than being unique to methodological decisions[36].

Discussions with our patient and public group highlighted the importance of being able to clearly explain why a patient is assigned to a cluster for healthcare applications. Without clearly separable clusters, this is difficult: in our case, we would be unable to assign an individual to a cluster without an external algorithm. Assignment algorithms, such as decision trees, may be used in this context to provide clear assignment thresholds, but would not be deterministic and would introduce new statistical error in assignment.

Clusters derived from EHR-BERT—incorporating sociodemographic characteristics—performed slightly better at predicting outcomes than those from Med-BERT, but these differences were largely eliminated once patient sociodemographic characteristics were included in the prediction model, indicating how EHR-BERT captures these relationships within the embeddings. Despite large variation in 1-year mortality and healthcare use across clusters (Figs. 2–4), differences were substantially attenuated after adjustment for sociodemographic characteristics, while relatively little of the overall variation in outcomes was explained by the clusters alone (Supplementary Table 1).

Furthermore, information on the clusters alone had lower predictive accuracy than individual-level data on the conditions or embeddings from which they were derived (Table 3). This reflects the expected information loss from reducing continuous patient representations into a small number of groups but highlights the practical challenge of balancing interpretability while preserving individual heterogeneity. The number of LTCs alone surprisingly performed better than cluster membership in predicting ED attendance and hospitalisation, demonstrating the importance of overall disease burden, and highlighting the value of considering simple disease counts alongside disease composition when developing approaches to anticipate future care needs.

Instead of clustering on sociodemographic characteristics and diseases, some studies have clustered on metrics of healthcare utilisation or outcomes[13,37]. While these clusters will reflect people with similar healthcare use, this may produce clusters that are dissimilar with respect to patient characteristics and disease patterns, making it difficult to identify effective interventions. Furthermore, being in a "high-utilisation" cluster now may not mean someone is in a high utilisation cluster in future[13]. Some segmentation approaches incorporate both demographics and outcomes, but tend to be derived by clinical expertise, rather than data-driven processes, such as assigning a patient to a high ED use segment based on a threshold of attendances[14,38]. Creating clusters combining both patient characteristics and outcomes using data-driven methods requires a weighting of the two. Extending our transformer architecture to fine-tune the patient embeddings to predict specific outcomes could allow for clusters that retain both the characteristics (from pre-training) and outcome associations (from fine-tuning). A recent data-driven approach developed by Huang et al. combines characteristics and outcomes, thus creating groups similar with respect to both, which is a promising avenue for future work[39].

Table 3 | Performance metrics for predicting ED attendance, hospital admission and mortality from different model inputs

Outcome	Additional variables	Count of LTCs AUC	Diseases AUC	Embedding AUC	8 clusters AUC
ED attendance	None	0.622	0.609	0.641	0.568
	Sociodemographics	0.642	0.639	0.646	0.609
	Sociodemographics and number of LTCs	N/A	0.639	0.647	0.628
Hospital admission	None	0.667	0.718	0.736	0.653
	Sociodemographics	0.714	0.730	0.739	0.692
	Sociodemographics and number of LTCs	N/A	0.730	0.739	0.710
Mortality*	None	0.656	0.831	0.865	0.669
	Sociodemographics	0.777	0.852	0.865	0.783
	Sociodemographics and number of LTCs	N/A	0.852	0.865	0.786

* Among patients ≥60 years

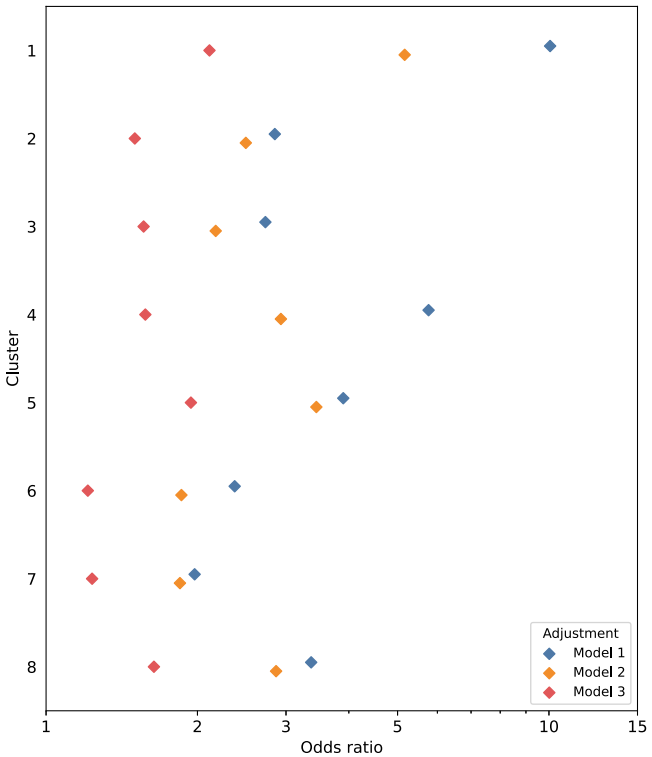


Fig. 3 | Odds ratios for any emergency hospital admission over one year for each cluster, compared to people with no LTCs. Model 1: unadjusted; Model 2: adjusted for sociodemographics; Model 3: adjusted for sociodemographics and number of LTCs per person. Vertical line at 1 represents OR for people with no LTCs. Point estimates and confidence intervals are shown in Supplementary Table 2. A comparison of outcomes among those with MLTC only is shown in Supplementary Fig. 14.

A strength of our study is the use of a large and nationally representative sample of patients from primary care in England[19]. We defined MLTC based on a broader set of LTCs than many studies, and examined a range of healthcare utilisation measures relevant to both primary and secondary care. Our embeddings were generated from a novel transformer algorithm developed in earlier work, which uses disease sequences, incorporating information on age, gender, ethnicity and deprivation[17]. We also tested a simpler embedding model which uses the disease sequences alone – direct comparison was challenging due to differences in the optimal number of clusters, but cardio-kidney-metabolic and lipid clusters were consistent across both methods, while explanatory and predictive performance was slightly lower. This suggests that the observed challenges in interpretability and predictive utility are not solely attributable to the inclusion of sociodemographic information and may reflect broader challenges of applying hard clustering to heterogeneous disease trajectories.

We applied k-means clustering, a widely used clustering algorithm which is computationally feasible with large datasets. However, k-means can generate unbalanced clusters with high-dimensional data. For this reason, we tested a reduction in the dimensionality of the data using PCA and did not find substantial differences in the resulting clusters. Applying a clustering model learned in one setting to a new population would require training of an additional predictive algorithm to learn assignment patterns. Although beyond the scope of this work, this is likely to be challenging to implement and would lead to greater loss of both the explanatory and predictive power of the clusters, representing a limitation to the real-world use of data-driven clusters of people in healthcare settings.

A limitation of generating sequence-based embeddings from EHR data is the capturing of patterns of healthcare interaction as well as disease biology, which may lead to clusters reflecting care processes rather than true phenotypes[40]. For example, cluster 4 included four common measures of elevated blood lipids, which are routinely measured together. We also assessed associations of clusters with outcomes over a relatively short 1-year time-frame; assessment over longer horizons in future research may help identify whether clusters capture meaningful longer-term differences in health. Trajectory-based approaches may be more adept at distinguishing variation in longitudinal patterns in outcomes, for example, stable versus escalating use of services.

Based on our findings, we propose the following considerations for future research using clustering approaches in MLTC:

What is the goal of clustering?

Clustering should be aligned to a clearly defined purpose, such as describing disease patterns, informing service design or risk stratification. In our study, clustering disease trajectories did not improve short-term risk prediction over individual-level disease representations. Where prediction is the primary objective, approaches retaining individual-level information or incorporating supervision may therefore be preferable[39], while unsupervised clustering may be better suited to generating interpretable groups for service or pathway design.

What is the unit of clustering?

Clustering people across trajectories of 212 LTCs resulted in substantial overlap between clusters with limited separation of outcomes. Clustering diseases may be more interpretable, as the number of conditions is far smaller, but requires an additional assignment step[10]. Alternatively, where the objective is to identify groups with distinct risks or outcomes,

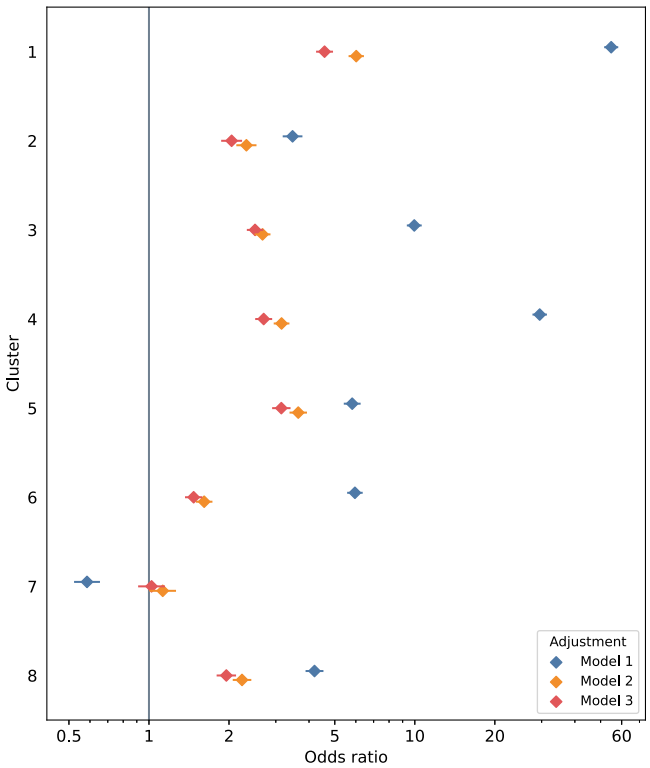


Fig. 4 | Odds ratios for death over one year for each cluster, compared to people with no LTCs. Model 1: unadjusted; Model 2: adjusted for sociodemographics; Model 3: adjusted for sociodemographics and number of LTCs per person. Vertical line at 1 represents OR for people with no LTCs. Point estimates and confidence intervals shown in Supplementary Table 2. A comparison of outcomes among those with MLTC only is shown in Supplementary Fig. 15.

segmentation approaches clustering on outcomes or utilisation patterns may be more appropriate.

What type of clustering approach should be used?

Hard, mutually-exclusive clustering may be unsuited to MLTC, where disease patterns commonly overlap, as observed here. Probabilistic or 'soft' clustering approaches allowing membership across multiple groups may better represent this heterogeneity[41,42].

Does clustering add value over simpler stratification?

In our analysis, cluster membership explained little variation in healthcare utilisation and mortality, providing less predictive information than individual conditions and in some cases less than simple disease counts. Clustering should therefore be benchmarked against simpler, interpretable approaches to determine the incremental value of additional complexity.

Should clusters be stratified across populations?

We observed large demographic differences in cluster membership. Prior stratification by characteristics such as age and sex may identify patterns driven less by these factors[43,44], but increases complexity of interpretation and reduces comparability across subgroups. Choice should therefore be guided by the intended application.

How should similarity and heterogeneity be evaluated?

Most variation in healthcare outcomes in our study occurred within, rather than between clusters. Evaluation should therefore assess within- and between-cluster heterogeneity in features or outcomes relevant to the intended purpose, rather than relying on statistical metrics of cluster separation alone.

How can generalisability be established?

Different embedding models produced different cluster configurations in our study, demonstrating sensitivity to modelling choices. Establishing generalisability therefore requires assessment of both stability across modelling assumptions and replicability across populations.

How will clusters be used in practice?

High within-cluster heterogeneity in our study limited the ability of cluster membership alone to identify individual care needs. Practical value is more likely where clusters align with recognised care pathways, support specific interventions and are reproducible across settings. Ultimately, implementation studies are needed to establish whether their use improves service delivery or patient outcomes.

Conclusion

We generated eight clusters of people with MLTC based on similar disease trajectories. Although clusters differed in sociodemographic profiles, disease composition, healthcare utilisation and mortality, interpretation was challenging due to substantial overlap in conditions across clusters and high residual heterogeneity in outcomes within clusters. Cluster membership also provided less predictive information than individual-level diseases, and for some outcomes, the number of LTCs alone, demonstrating the information lost when heterogeneous trajectories are reduced to discrete groups. Future research should focus on approaches that better balance simplification and interpretability with preserving individual heterogeneity and evaluate whether clustering can support practical improvements in patient care and clinical outcomes.

Data availability

This study used data from the Clinical Practice Research Datalink (CPRD) which are not publicly available, but may be requested from CPRD, subject to approvals. More information is available at: https://www.cprd.com/access-data

Code availability

Code lists developed during this project are available in a dedicated project repository[45].

Received: 21 January 2026; Accepted: 8 September 2026;
Published online: 21 September 2026

References

1. Johnston, M. C., Crilly, M., Black, C., Prescott, G. J. & Mercer, S. W. Defining and measuring multimorbidity: a systematic review of systematic reviews. *Eur. J. Public Health* **29**, 182–189 (2019).
2. Pearson-Stuttard, J., Ezzati, M. & Gregg, E. W. Multimorbidity—a defining challenge for health systems. *Lancet Public Health* **4**, e599–e600 (2019).
3. Nunes, B. P., Flores, T. R., Mielke, G. I., Thumé, E. & Facchini, L. A. Multimorbidity and mortality in older adults: a systematic review and meta-analysis. *Arch. Gerontol. Geriatr.* **67**, 130–138 (2016).
4. Soley-Bori, M. et al. Impact of multimorbidity on healthcare costs and utilisation: a systematic review of the UK literature. *Br. J. Gen. Pract.* **71**, e39–e46 (2021).
5. Makovski, T. T., Schmitz, S., Zeegers, M. P., Stranges, S. & van den Akker, M. Multimorbidity and quality of life: systematic literature review and meta-analysis. *Ageing Res. Rev.* **53**, 100903 (2019).
6. The taskforce on multiple conditions. *Just One Thing after Another: Living with Multiple Conditions*. https://www.richmondgroupofcharities.org.uk/wp-content/uploads/2018/10/final_just_one_thing_after_another_report_-_singles.pdf https://doi.org/10.1007/s11016-010-9485-1 (2018).
7. Stokes, J., Guthrie, B., Mercer, S. W., Rice, N. & Sutton, M. Multimorbidity combinations, costs of hospital care and potentially

preventable emergency admissions in England: a cohort study. *PLoS Med.* **18**, e1003514–e1003514 (2021).
8. Busija, L., Lim, K., Szoeke, C., Sanders, K. M. & McCabe, M. P. Do replicable profiles of multimorbidity exist? Systematic review and synthesis. *Eur. J. Epidemiol.* **34**, 1025–1053 (2019).
9. Ferris, J. K. et al. A systematic review and meta-analysis of disease clusters in multimorbidity. *Nat. Commun.* **17**, 634 (2025).
10. Beaney, T. et al. Assigning disease clusters to people: a cohort study of the implications for understanding health outcomes in people with multiple long-term conditions. *J. Multimorbidity Comorbidity* **14**, 26335565241247430 (2024).
11. Robertson, L. et al. Identifying multimorbidity clusters in an unselected population of hospitalised patients. *Sci. Rep.* **12**, 5134 (2022).
12. Smith, A. et al. Identifying clusters of people with multiple long-term conditions using large language models: a population-based study. *npj Digit. Med.* **8**, 453 (2025).
13. Beaney, T. et al. Patterns of healthcare utilisation in children and young people: a retrospective cohort study using routinely collected healthcare data in Northwest London. *BMJ Open* **11**, e050847 (2021).
14. Yan, S., Kwan, Y. H., Tan, C. S., Thumboo, J. & Low, L. L. A systematic review of the clinical application of data-driven population segmentation analysis. *BMC Med. Res. Methodol.* **18**, 121 (2018).
15. Zhou, Y. Y., Wong, W. & Li, H. Improving care for older adults: a model to segment the senior population. *Perm. J.* **18**, 18–21 (2014).
16. Torkki, P., Leskelä, R.-L., Mustonen, P., Linna, M. & Lillrank, P. How to extend value-based healthcare to population-based healthcare systems? Defining an outcome-based segmentation model for health authority. *BMJ Open* **13**, e077250 (2023).
17. Beaney, T. et al. Comparing natural language processing representations of coded disease sequences for prediction in electronic health records. *J. Am. Med. Inform. Assoc.* ocae091 https://doi.org/10.1093/jamia/ocae091 (2024).
18. Shmatko, A. et al. Learning the natural history of human disease with generative transformers. *Nature* **647**, 248–256 (2025).
19. Wolf, A. et al. Data resource profile: clinical practice research datalink (CPRD) Aurum. *Int. J. Epidemiol.* **48**, 1740–1740g (2019).
20. Herrett, E. et al. Data resource profile: clinical practice research datalink (CPRD). *Int. J. Epidemiol.* **44**, 827–836 (2015).
21. Lewis, J. D., Bilker, W. B., Weinstein, R. B. & Strom, B. L. The relationship between time since registration and measured incidence rates in the General Practice Research Database. *Pharmacoepidemiol. Drug Saf.* **14**, 443–451 (2005).
22. Ministry of Housing & Communities & Local Government. English indices of deprivation 2019 https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 (2019).
23. Delmestri, A. & Prieto-Alhambra, D. CPRD GOLD and linked ONS mortality records: Reconciling guidelines. *Int J. Med. Inf.* **136**, 104038 (2020).
24. Kuan, V. et al. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *Lancet Digit. Health* **1**, e63–e77 (2019).
25. Head, A. et al. Inequalities in incident and prevalent multimorbidity in England, 2004–19: a population-based, descriptive study. *Lancet Healthy Longev.* **2**, e489–e497 (2021).
26. Beaney, T. et al. Effect of timeframes to define long term conditions and sociodemographic factors on prevalence of multimorbidity using disease code frequency in primary care electronic health records: retrospective study. *BMJ Med.* **3**, 474 (2024).
27. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. in *Proceedings of NAACL-HLT 2019* vol. 1 4171–4186 (Association for Computational Linguistics, 2019).
28. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digit. Med.* **4**, 1–13 (2021).
29. Beaney, T. et al. Identifying multi-resolution clusters of diseases in ten million patients with multimorbidity in primary care in England. *Commun. Med.* **4**, 102 (2024).
30. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
31. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M. & Perona, I. An extensive comparative study of cluster validity indices. *Pattern Recognit.* **46**, 243–256 (2013).
32. Steinbach, M., Ertöz, L. & Kumar, V. The challenges of clustering high dimensional data. In *New Directions in Statistical Physics* (ed. Wille, L. T.) 273–309 https://doi.org/10.1007/978-3-662-08968-2_16 (Springer, 2004).
33. Mueller, A. Wordcloud. https://github.com/amueller/word_cloud/blob/main/CITATION.cff (2023).
34. StataCorp. *Stata 17 Base Reference Manual*. (Stata Press, 2021).
35. Ndumele, C. E. et al. Cardiovascular-kidney-metabolic health: a presidential advisory from the american heart association. *Circulation* **148**, 1606–1635 (2023).
36. Beaney, T. et al. Clustering in multiple long-term conditions: methodological and translational challenges and solutions. *J. R. Soc. Med.* 1410768251395387 https://doi.org/10.1177/01410768251395387 (2025).
37. Vuik, S. I., Mayer, E. & Darzi, A. A quantitative evidence base for population health: applying utilization-based cluster analysis to segment a patient population. *Popul. Health Metr.* **14**, 44 (2016).
38. Chong, J. L., Lim, K. K. & Matchar, D. B. Population segmentation based on healthcare needs: a systematic review. *Syst. Rev.* **8**, 202 (2019).
39. Huang, Y. et al. Deep significance clustering: a novel approach for identifying risk-stratified and predictive patient subgroups. *J. Am. Med Inf. Assoc.* **28**, 2641–2653 (2021).
40. Beaney, T. et al. Identifying potential biases in code sequences in primary care electronic healthcare records: a retrospective cohort study of the determinants of code frequency. *BMJ Open* **13**, e072884 (2023).
41. Violán, C. et al. Soft clustering using real-world data for the identification of multimorbidity patterns in an elderly population: cross-sectional study in a Mediterranean population. *BMJ Open* **9**, e029594 (2019).
42. Marengoni, A. et al. Patterns of multimorbidity in a population-based cohort of older people: sociodemographic, lifestyle, clinical, and functional differences. *J. Gerontology: Ser. A* **75**, 798–805 (2020).
43. Smith, A. et al. Identifying clusters of people with multiple long-term conditions using large language models: a population-based study. 2025.02.14.25322277 Preprint at https://doi.org/10.1101/2025.02.14.25322277 (2025).
44. Krauth, S. J. et al. Association of latent class analysis-derived multimorbidity clusters with adverse health outcomes in patients with multiple long-term conditions: comparative results across three UK cohorts. *eClinicalMedicine* **74**, 102703 (2024).
45. Beaney, T. Clustering multiple long-term conditions https://tbeaney.github.io/MMclustering/. Accessed 02/09/2026

Acknowledgements

This research is funded through a clinical PhD fellowship awarded to TB from the Wellcome Trust 4i programme at Imperial College London and a postdoctoral clinical fellowship from the NIHR Imperial Biomedical Research Centre. JC acknowledges support from the Wellcome Trust (215938/Z/19/Z). TW, AM and PA acknowledge support from the National Institute for Health and Care Research (NIHR) Applied Research Collaboration Northwest London. MB acknowledges support from EPSRC grant EP/N014529/1 supporting the EPSRC Centre for Mathematics of Precision

Healthcare. Infrastructure support was provided by the NIHR Imperial Biomedical Research Centre (BRC). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR, the Wellcome Trust or the Department of Health and Social Care. Data management was provided by the Big Data and Analytical Unit (BDAU) at the Institute of Global Health Innovation (IGHI), Imperial College London.

Author contributions

T.B. conceptualised the study, conducted the data management and formal analysis, and wrote the first draft of the manuscript. J.C., T.W., A.M., M.B., and P.A. contributed to the study design, methods, supervision, interpretation of findings, and reviewing and editing the manuscript.

Competing interests

The authors have no competing interests to declare. Thomas Beaney is an Editorial Board member for Communications Medicine but was not involved in the editorial review or peer review, nor the decision to publish this article.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-026-01917-8.

Correspondence and requests for materials should be addressed to Thomas Beaney.

Peer review information Communications Medicine thanks Maria Pikoula, and Elia Yathie Matsumoto for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.Consent to participateIndividual patient consent was not required, as the study used pseudonymised routinely collected healthcare data provided by CPRD. General practices consent to contribute patient data to CPRD and patients are able to opt out of data-sharing for research purposes.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.

Thomas Beaney[1,2] ✉, **Jonathan Clarke**[3], **Thomas Woodcock**[2], **Azeem Majeed**[2], **Mauricio Barahona**[3] & **Paul Aylin**[2]

[1]The George Institute for Global Health, Imperial College London, London, UK. [2]Department of Primary Care and Public Health, Imperial College London, London, UK. [3]Centre for Mathematics of Precision Healthcare, Department of Mathematics, Imperial College London, London, UK. ✉e-mail: thomas.beaney@imperial.ac.uk