Current Epidemiology Reports (2019) 6:364–372
https://doi.org/10.1007/s40471-019-00211-7

PHARMACOEPIDEMIOLOGY (U HAUG, SECTION EDITOR)

Machine Learning for Feature Selection and Cluster Analysis in Drug Utilisation Research

Sara Khalid[1] · Daniel Prieto-Alhambra[1]

Published online: 27 July 2019

Abstract

Purpose of Review Machine learning methods are increasingly used in health data mining. We describe current unsupervised learning methods for phenotyping and discovery and illustrate their application for detecting features and sub-groups related to drug use within a population.

Recent Findings Patient representation or phenotyping and discovery is one of the main branches of health data analysis. Phenotyping concerns identifying features that are representative of the population from raw patient data. Discovery involves analysing these features, for example, to identify patterns in the population such as sub-groups and to predict outcomes. Most studies use unsupervised learning methods for phenotyping as they are suited for data-driven feature extraction. We describe some of the commonly used methods and demonstrate their use in feature selection followed by cluster analysis.

Summary Unsupervised learning methods can be used to extract the features of and identify sub-groups within specific populations. We demonstrate the potential of these methods and highlight the associated challenges, which researchers may find useful in understanding the suitability of these methods for analysing health data.

Keywords Phenotyping and discovery · Electronic health records · Unsupervised learning · Feature selection · Cluster analysis · Autoencoder

Introduction

Drug utilisation research traditionally involves characterising pre-defined patient features, usually related to the indications and contraindications of a drug or drug class for regulatory purposes and post-marketing surveillance. This approach is hypothesis-driven and can potentially fail to identify particular sub-groups of drug users that might be of interest for further studies on the risk-benefit of the drugs.

For some time, it has been possible to use information extracted from electronic health records (EHRs) and claims data for drug utilisation research. EHRs are data-rich and, crucially, represent real-world use of the drugs in question within the community. In parallel, recent advances in data science have led to the development of techniques that can be used to mine these, large-scale data and allow data-driven analysis. This approach has the potential to capture patterns that may not have been considered in a hypothesis-driven setting and therefore reveal more about the actual use of drugs in the real world.

EHR mining has been studied widely [1, 2], and a range of techniques are used, including data mining [3], natural language processing [4] and, more recently, deep learning [5••]. Most studies use a two-step approach of phenotyping and discovery [6], shown in Fig. 1.

Phenotyping and Discovery

EHR data can be complex in terms of both volume (number of patient records) and variety (types of information stored). A typical EHR database can include patient information such as clinical diagnoses codes, demographic data, laboratory and imaging results, vital signs and free-text notes. The first step in analysing EHR data is phenotyping, which involves selecting clinically relevant features from the raw data (Fig. 1). Once a set of features has been extracted, knowledge

This article is part of the Topical Collection on *Pharmacoepidemiology*

✉ Sara Khalid
sara.khalid@ndorms.ox.ac.uk

[1] Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Old Road, Oxford OX37LD, UK

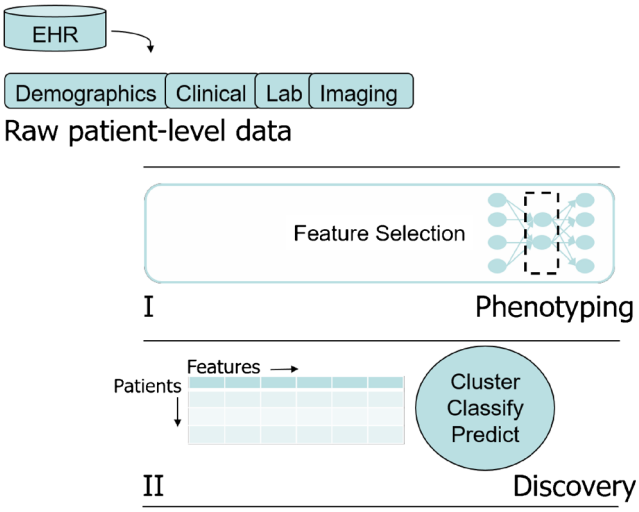

Fig. 1 The process of analysis of electronic health record (EHR) data for phenotyping and discovery. In the first stage, raw patient data are analysed to extract meaningful features. These features are used to build models (e.g. cluster analysis, classification and prediction models) in the second stage.

discovery can follow. The features are analysed at a population-specific or patient-specific level to determine if there are sub-groups (cluster analysis) or if they can be used to derive new diagnostic criteria (classification) or to determine prognosis (prediction).

Machine learning methods have recently been used in EHR mining [5••, 7]. In particular, unsupervised learning methods have shown promise for feature selection [8••, 9••, 10] before clustering and prediction tasks.

In this paper, we explain feature selection and cluster analysis in the context of EHR-based drug utilisation research. We illustrate the methods using a case study on anti-osteoporosis drugs.

Anti-osteoporosis drugs are commonly prescribed as preventative therapies to patients at risk of fragility fractures [11]. The case study is based on data from the SIDIAP database (www.sidiap.com), which provides anonymised electronic general practice-level data from Catalonia (Spain) [12].

As a first step, it is important to characterise the study population with respect to key variables that are relevant in the clinical context. In our case study, which included 37,996 patients using anti-osteoporosis drugs, 32 variables were extracted for each patient, including clinical diagnosis codes, lab tests and demographics. The characteristics are summarised in Table 1.

Before any analysis, data can be examined for obvious correlations in the structure. In our example, an initial assessment of the data did not suggest strong correlations (where a strong correlation would have Pearson's correlation > 0.6 for any variable pair). Another statistic used for assessing grouping structure is Hopkin's statistic [13], and a value 0.44 for it indicated that the data were distributed in a random manner.

Feature selection and cluster analysis

When performing cluster analysis, an underlying assumption is that there are sub-groups or clusters within the population (in our example, "sub-groups of anti-osteoporosis drug users). The "true" number and nature of these clusters are unknown, which make this an unsupervised learning problem. Therefore, expert knowledge is often used as a basis for determining the number of clusters (k) to be derived. We also do not know which of the variables (and combinations thereof) are most characteristic of the population and can thus be considered as features. Our task is to first identify the most characteristic variables of the given population (feature selection) and then to learn the structure of the k clusters based on these features (cluster analysis).

Feature Selection

The autoencoder [14] is an unsupervised learning algorithm for feature selection using unlabelled data. It is a feedforward neural network. In its simplest form, its architecture comprises an input layer that feeds into a hidden layer, which in turn feeds into an output layer. Consider a D-dimensional dataset $X = \{x_1, x_2, \ldots x_D\}$, where D is the number of variables, presented at the input layer. The autoencoder attempts to reconstruct X at the output layer. In other words, it models the identity function $f(x) = x$ [15]. To do so, the hidden layer is forced to learn a compressed, weighted representation of the data X presented at the input layer, which is then reconstructed at the output layer as $\hat{X}$. The autoencoder is suitable for tasks such as dimensionality reduction and feature selection because it produces this compressed representation of the data.

The learning process depends on the architecture of the autoencoder (the number of nodes in the hidden layer) and the sparsity parameter[1] ρ, which enables the compressed representation. The optimal architecture is one for which these two parameters result in the smallest reconstruction error (RMSE) between X and $\hat{X}$.

w_{dj}, the weight assigned to the dth variable at the jth node of the hidden layer, can be used to generate a measure of the "importance" of that variable in the reconstruction of the dataset, where $d \in \mathbb{Z}$: $1 \leq d \leq D$ and $j \in \mathbb{Z}$: $1 \leq j \leq J$.

It is not necessarily straightforward to interpret the expression combining the weights at the hidden layer. In our example, we took a simple approach. The weight of a variable d at a given node j signified its importance in activating that node. The greater the weight of a variable, the more important it was for the activation. We therefore considered the average weight of a variable across all of the J nodes, $\hat{w}_{dj} = \sum_{j=1}^{J} w_{dj}$. A

[1] The sparsity parameter ρ imposes a constraint on the activation of the hidden units, which reduces the dependency between features. $D >> N$.

Table 1 Prevalence of features in the dataset*

Variables	Description	Prevalence (%)	Variables	Description	Prevalence (%)
Elderly	Age ≥ 60 years	83.3	Chronic kidney failure (CKF) test	Chronic kidney failure diagnosed from lab results in the last 2 years	10.29
Female	Female gender	79.4	History of haemorrhagic stroke	Haemorrhagic stroke from ICD10 codes	0.28
Obesity	BMI ≥ 30	33.6	Pulmonary embolism	History of pulmonary embolism	0.55
Smoking	Current use[a]	7.7	Deep vein thrombosis	History of deep vein thrombosis	0.85
Drinking	Current use[a]	1.1	Temporary ischaemic attack (TIA)	History of temporary ischaemic stroke/attack	1.07
Comorbid	Charlson index ≥ 2	52.1	Myocardial infarction (MI)	History of myocardial infarction	1.91
Steroid user	Current use[a]	13.3	Gynaecology	Referral to gynaecology or woman care in previous 365 days	3.00
Sedative user	Current use[a]	46.3	Stroke	History of stroke	4.09
Previously fractured	Any major fracture (including hip and non-hip)	18.5	Ischaemic heart disease (IHD)	History of ischaemic heart disease	5.35
Contraceptive or HRT user	Current use[a]	6.3	Rheumatology	Referral to rheumatology in previous 365 days	8.00
Anti-coagulant user	Current use[a]	6.7	Cancer	History of cancer of any type	8.17
Aromatase Inhibitor user	Current use[a]	2.2	Varicose veins	History of varicose veins	15.00
Hip osteonecrosis	History of osteonecrosis of hip	0.00	Orthopaedics	Referral to orthopaedics or trauma in the last 365 days	16.83
Osteonecrosis	History of osteonecrosis	0.21	Type 2 diabetes	History of type 2 diabetes	21.45
Nephrotic syndrome	History of nephrotic syndrome	0.05	Nurse visits	At least one primary care nurse appointment in the last 365 days	95.17
Chronic kidney failure (CKF)	History of chronic kidney failure, from ICD10 codes	5.44	GP visits	At least one general practitioner appointment in the last 365 days	99.20

[a] As opposed to previous use or never used

*Data were extracted for years 2012–2016. 107,240 patients were found to have missing data for BMI, smoking and drinking and were excluded from this case study, although we recommend multiple imputations prior to analysis for a complete study

Reprinted, with permission, from IEEE, Cluster Analysis to Detect Patterns of Drug Use from Routinely Collected Medical Data, June 1, 2018

BMI, body mass index; *HRT*, hormone replacement therapy

variable with a low $\hat{w}_{dj}$ would have less importance than a variable with a higher $\hat{w}_{dj}$. A selection threshold can be defined such that variables with weights above the threshold are considered selected features.

In our example, a two-layer autoencoder was constructed with D-dimensional input and output layers and one hidden layer containing J nodes. To identify the number of hidden nodes, J, and sparsity parameter, ρ, of the optimal autoencoder, we performed a grid search using $1 < J < 20$ and $0.001 < \rho < 0.995$. For the optimal model, $\hat{w}_{dj}$ was estimated for each variable. Variables with weights $\hat{w}_{dj} > 0.5$ at any node in the hidden layer were considered features of the dataset.

It is useful to apply feature selection to different subsets of the data to assess whether the results differ with the number of input variables. In our example, we applied feature selection to a subset of the data containing 12 variables that are considered risk factors in the osteoporosis literature and to the full dataset containing 32 variables.

Evaluating Feature Selection

As feature selection in unsupervised learning is purely data driven, it is often compared with other statistical approaches, and expert opinion is typically sought to evaluate the results. In our example, features selected by the autoencoder were compared with those obtained using principal component analysis (PCA). PCA is another method commonly used for feature selection and dimensionality reduction. It selects the variables that best explain the variability in the data.

Independently, we polled experts to ask which variables they believed were taken into consideration by general practitioners when assessing someone's risk of needing treatment for osteoporosis, i.e. were risk factors. The poll identified nine variables: age, female gender, obesity, smoking, alcohol use, comorbidity, steroid use, sedative use and fracture history. The expert-identified risk factors were used as the reference against which the features selected by the autoencoder were compared.

In the subsequent discovery stage, variables selected as features by the autoencoder were used in cluster analysis to derive sub-groups of anti-osteoporosis drug users.

Cluster Analysis

There are several algorithms for cluster analysis [16], and *k*-means [17] and hierarchical [18] clustering are two of the most well-known and commonly used ones. They produce "hard" clustering, where exactly one cluster is assigned to a participant. It is also possible to produce "soft" clustering, where a participant may belong to more than one clusters, depending on the degree of membership. Gaussian mixture models and fuzzy *c*-means clustering can perform "soft" clustering [19]; however, this approach is often not applicable for data that are binary in nature.

Hierarchical Clustering

Recalling the dataset $X = \{x_1, x_2, \ldots x_D\}$, we may consider the *i*th participant to be represented in the *D*-dimensional data space by X_i, where $i \in \mathbb{Z}$: $1 \leq i \leq n$ and n is the number of participants in the dataset. First, each participant is considered to be a cluster, so that there are n clusters to begin with. Next, the closest clusters are merged, using a measure of closeness (e.g., the Euclidean distance in *D*-dimensional space). This process of merging carries on until all of the participants have been merged into either a pre-specified number of clusters (k) or one cluster[2].

k-means Clustering

First, the number of clusters k has to be pre-specified. Then k randomly selected points in the *D*-dimensional space are initialised as cluster centroids. Next, a participant is assigned to the cluster centroid that they are closest to. As with hierarchical clustering, closeness may be calculated using, e.g. Euclidean distance (other types of distance measures are also used, e.g. Hamming, Mahalanobis, city-block, etc.). The position of a cluster centroid is re-calculated based on the positions of the participants assigned to it. Participants continue to be re-assigned and centroids re-calculated until there are no further changes in the positions of any of the k clusters.

[2] This is an example of agglomerative clustering. Divisive clustering can also be performed, in which case the participants divided into clusters instead of being merged.

Internal Cluster Evaluation

The purpose of evaluation is to judge, in an objective way, how well the clustered data fit within the candidate k clusters. This can be done by checking (a) if individual clusters are homogenous and (b) if they are well-separated from the other clusters. The optimal number of clusters $\hat{k}$ is found if these criteria are met.

In the case study, in search of $\hat{k}$, hierarchical and *k*-means clustering were performed on the dataset using the features selected by the autoencoder model. Based on expert opinion, it was reasonable to derive as many as 10 sub-groups of anti-osteoporosis drug users. Nevertheless, we set the candidate number of clusters from $k = 2$ through to $k = 20$. For each value of k, the clustering process was carried out 100 times, using a random sample of 1000 participants each time. To measure the closeness of points in the data space, we used the squared Euclidean and city-block distance measures. For evaluation of the resultant clusters, we used commonly used criteria [20, 21], including silhouette, Calinski-Harabasz (CH) and gap.

External Cluster Evaluation

Internal evaluation alone is often not sufficient for evaluating clusters. External evaluation can be performed using information about the population that was not used as a feature in the cluster model. In our example, information on bone mineral density and incident hip fracture risk was available in the SIDIAP database. These are key proxies of osteoporosis risk and indication for anti-osteoporosis drug therapy. Since bone mineral density and hip fracture risk were not used in the model generation process, we could use them for evaluating the resulting clusters.

$$\text{Fracture risk was estimated as } \frac{\text{number of fractures since start of study}}{\text{follow-up time, totaled for all persons}} \times 1000 \text{ in units of 1000 person years (py).}$$

The bone mineral density and fracture risk of the k clusters were examined.

Presenting the Results of Feature Selection and Cluster Analysis

The results of feature selection should ideally reflect how the model(s) used compare with expert opinion. The results of cluster analysis should not only be assessed in light of internal

and external evaluation, but also by considering whether they are clinically plausible. We illustrate this using the results of our case study.

Feature Selection

The minimum error between the original and reconstructed datasets (RMSE = 0.08) was obtained when the number of hidden nodes in the autoencoder was set to $J = 5$ and the sparsity parameter was set to $\rho = 0.5$.

Selecting From 12 Variables

When selecting from the subset of 12 variables, the autoencoder assigned high weights ($\hat{w}_{dj} > 0.5$) to 8 of the 12 variables (Fig. 2a). These 8 variables were the same variables independently identified by clinical experts as risk factors. The only clinically identified risk factor not ranked highly by the autoencoder was alcohol consumption.

Selecting From 32 Variables

When selecting from all 32 variables, 7 of the 8 variables were selected again. Steroid use was not ranked highly (Fig. 2b). An additional 3 variables were selected: varicose veins, type 2 diabetes and cancer.

When selecting from both the 12-variable subset and the 32-variable full dataset, the ranking of the selected variables did not exactly match their prevalence in the dataset. For instance, obesity was the highest-ranking variable in the selection although it was present in 34% of the population, whereas being female was the seventh highest-ranking variable. However, in both scenarios, the features with the highest prevalence in the dataset were all selected.

Comparison With PCA

PCA's ranking of the variables explained 19% of the variation in the dataset. When selecting from the 12-variable subset, the

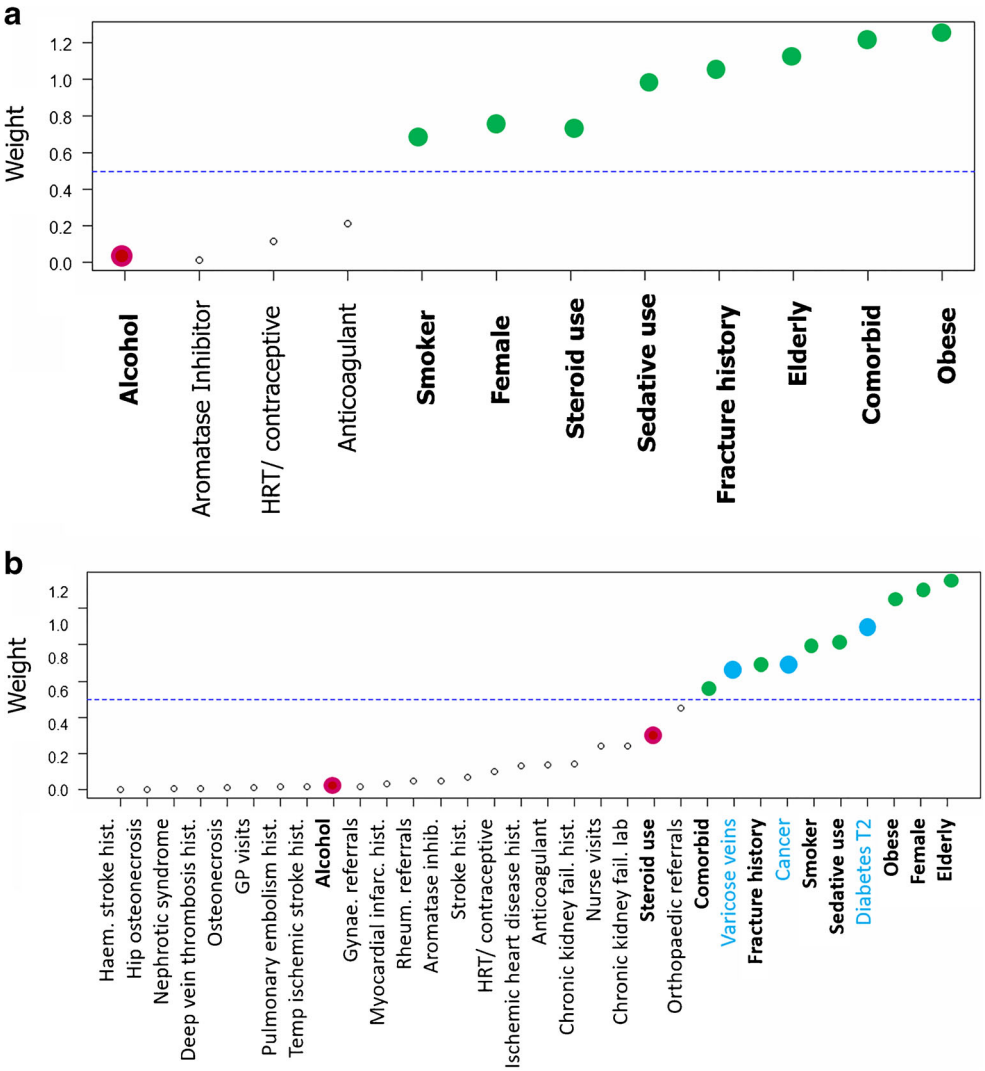

Fig. 2 Weights assigned by the autoencoder to the variables in the dataset containing **a** 12 and **b** 32 risk factors. The dashed line represents a threshold of $\hat{w}_{dj} > 0.5$. A green circle indicates a feature selected by both expert opinion and the autoencoder. A red circle indicates a feature selected by expert opinion but not by the autoencoder. A blue circle indicates a feature not selected by the autoencoder in the 12-variable dataset but selected in the 32-variable dataset

variable ranking by the autoencoder agreed with the variable ranking by the PCA method, with some exceptions. Obesity was ranked first by the autoencoder and second by PCA, and vice versa for comorbidity. Fracture history was ranked fourth by the autoencoder and seventh by PCA. Being female was ranked seventh by the autoencoder and fourth by PCA.

There was less agreement between PCA and the autoencoder when selecting from the 32-variable dataset.

Cluster Analysis and Evaluation

Number of Clusters

Figure 3 shows the results for *k*-means and hierarchical clustering applied to the features selected by the autoencoder model (with alcohol consumption added). In general, higher CH, gap and silhouette values indicate better within-cluster homogeneity and between-cluster separation.

For both *k*-means and hierarchical clustering, CH decreased from $k = 2$ onward, suggesting $\hat{k} = 2$. The smallest gap resulted when $k = 2$ for both *k*-means and hierarchical clustering and, as *k* increased, no clear elbow point was found to show where the gap was maximised. The silhouette criterion showed similar results to the gap criterion for *k*-means (indicating better clustering as *k* increased). However, for hierarchical clustering, it suggested that the clustering solution initially became worse as *k* increased from $k = 2$ to $k = 4$, stabilised from $k = 5$ and became worse again at $k = 8$. These results were based on Euclidean distance. Similar results were obtained using the city-block metric (data not shown).

Figure 3 demonstrates that the hierarchical and *k*-means clustering did not necessarily agree in their clustering solution. It also shows that there was no clear elbow point or stable state on the evaluation curves. The error search methods [22] that are typically used to determine $\hat{k}$ were therefore not appropriate here. We thus cannot conclude that there was an obvious optimal number of clusters in our data. This highlights an important complexity related to cluster analysis: it is not always straightforward to determine an optimal grouping distribution. We noted this as an unsurprising complexity due to the challenging nature of our real-world EHR data and continued to examine the structure of the clusters.

Cluster Structure

Figure 4 shows the composition of the clusters obtained using hierarchical clustering. At $k = 2$, the main feature that seemed to distinguish the two clusters was gender. With increasing *k*, the predominantly female cluster was divided, whereas the male cluster remained as it was. The figure demonstrates the changing composition of the clusters as they are further divided, up to $k = 5$. Table 2 summarises the characteristics of the

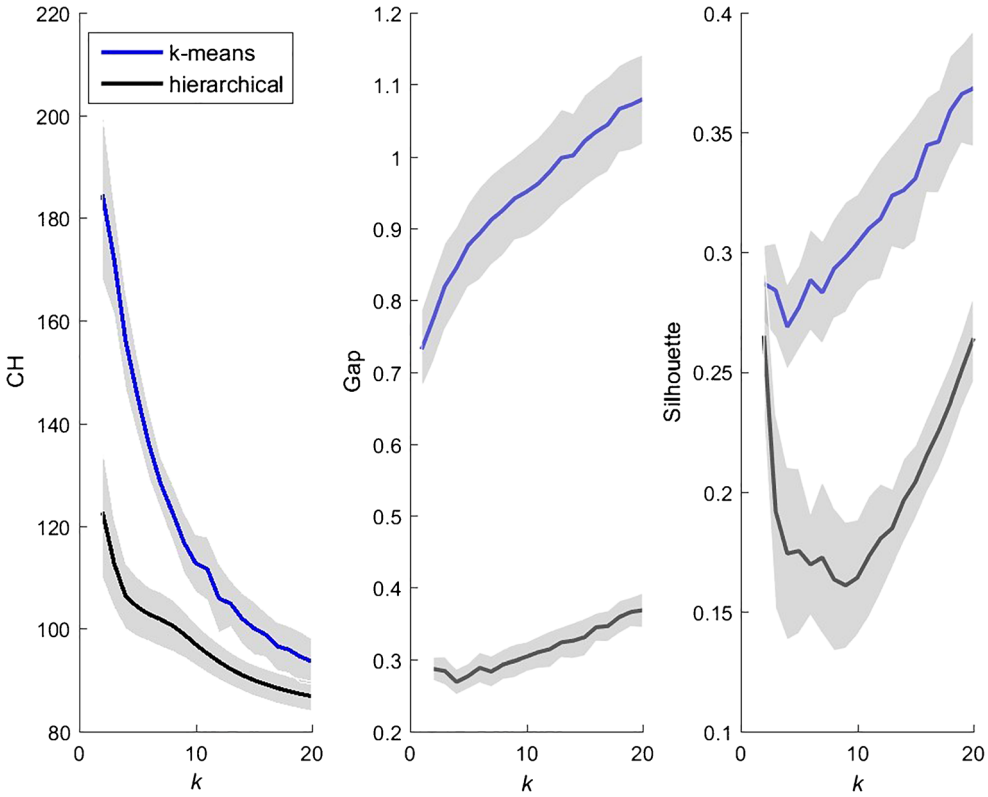

Fig. 3 Internal evaluation of hierarchical (black) and *k*-means (blue) clustering solutions using the CH (Calinski-Harabasz) (left), gap (centre) and silhouette (right) criteria. Error bars (grey) show the standard deviation over 100 iterations. (Reprinted, with permission, from IEEE, Cluster Analysis to Detect Patterns of Drug Use from Routinely Collected Medical Data, June 1, 2018)

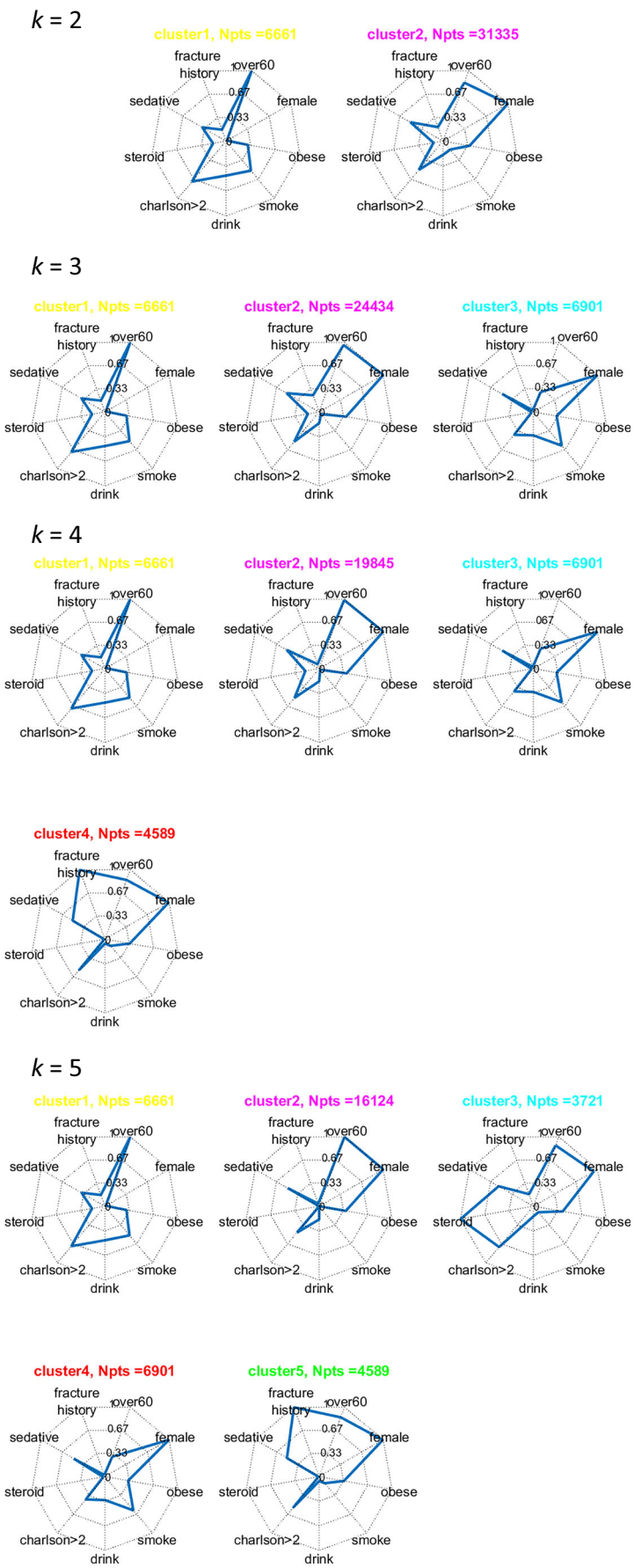

Fig. 4 Distribution of the risk factors within a cluster obtained using hierarchical clustering. A feature can take values between 0 and 1, where, e.g. female = 0 indicates all of the participants in this cluster are male and female = 1 indicates all of the participants are female. The clusters corresponding to $k = 2$, $k = 3$, $k = 4$ and $k = 5$ are shown in the panels from top to bottom, respectively. The label "charlson>2" corresponds to the risk factor "comorbid" and the label "over60" corresponds to the risk factor "elderly". (Reprinted, with permission, from IEEE, Cluster Analysis to Detect Patterns of Drug Use from Routinely Collected Medical Data, June 1, 2018)

Table 2 Prevalence of features in each of the five clusters detected using hierarchical clustering, where the minimum and maximum prevalence can be 0 and 1, respectively

Feature	Prevalence				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Elderly	1.00	1.00	0.88	0.25	0.86
Female	0.01	1.00	0.94	0.86	0.99
Obese	0.29	0.36	0.40	0.27	0.34
Smoker	0.51	0	0.11	0.51	0.12
Drinker	0.46	0.18	0.14	0.27	0.06
Comorbid	0.71	0.47	0.72	0.35	0.55
Steroid user	0.18	0	1.00	0.02	0.01
Sedative user	0.37	0.49	0.54	0.42	0.50
Fracture history	0.16	0.04	0.17	0.02	1.00

Reprinted, with permission, from IEEE, Cluster Analysis to Detect Patterns of Drug Use from Routinely Collected Medical Data, June 1, 2018.

five clusters. This characterisation often helps researchers to assign "names" to the derived clusters and to further interpret them, e.g. in our example, cluster 1 could be referred to as the male cluster, cluster 5 could be the elderly women with fracture history, whereas cluster 4 could be referred to as younger women with no fracture history.

For the external evaluation, the fracture risk of the cluster of elderly women with fracture history (cluster 5) was, as expected, the highest (10.5/1000py) and their hip bone mineral density was the lowest (T score[3] = − 2.2). Clusters 1, 2 and 3 had a fracture risk of 4, 6 and 6.5 per 1000 py, respectively. Cluster 4 had a lower fracture risk (1.5/1000 py) than even the general source population (2.23/1000 py) [23]. This also appears plausible since this cluster comprised younger people with no previous fractures. And their hip bone mineral density was higher than that of cluster 5 (T score = − 1.6). However, despite having a "healthy" hip T score, cluster 4 participants had an average spine bone mineral density that was low enough to be osteoporotic (T score = − 2.7), which may explain why this seemingly healthy cluster was prescribed anti-osteoporosis drugs. In this manner, external evaluation can aid in the interpretation of the derived clusters and in judging their clinical plausibility.

Discussion and Conclusion

In this review article, we have explained and demonstrated the use of unsupervised machine learning methods for feature selection and cluster analysis of real-world EHR data for drug utilisation research. Although the case study presented had a limited number of variables, our intent was to show how the methods perform for both small and larger numbers of variables. A consistent set of features was selected regardless of the number of variables entered into the model.

Feature selection and cluster analysis are difficult to assess in the absence of a gold standard. In the case study, we were able to compare the selected features with clinical expert opinion. In reality, it might not always be possible to have a reference to compare against, which is precisely what makes the learning task unsupervised in the first place.

Internal evaluation of the results of clustering the dataset using the selected features exposed the difficulty in deriving an optimal number of clusters. It showed that extracting patterns from real-world data with complex underlying structures may require examining the clusters using information not included in the clustering model. Examining the bone mineral density and fracture risks for the detected groups, for instance, aided in understanding the structures of the clusters and demonstrated how cluster analysis can help to develop and characterise sub-group profiles.

When interpreting the results, it is important to consider whether the analysis was based on complete data, and if not, how any missing data were handled. In the example presented here, the derived clusters only represent the sub-population of anti-osteoporosis drug users who reported their data in full.

Another constraint when analysing EHR data is that some clinical variables are often recorded in a dichotomised fashion, e.g. smoking status is recorded as "yes" or "no". If all variables are binary, the choice of cluster analysis methods to be used can become limited.

A final note on why one might conduct such an analysis. As demonstrated, feature selection can provide a starting point if it is not known a priori which of many features should be chosen. Once reasonable features are selected, they can be used for discovery analysis, such as detecting clinically plausible subgroups. In this manner, phenotyping and discovery can help us to discover new or hidden drug use patterns or sub-groups.

Despite interest in machine learning methods, their uptake in drug utilisation research has been slow. It is hoped that by considering the strengths and limitations showcased here, researchers will be better positioned to make informed decisions on the suitability of these methods for tasks such as sub-group detection.

Acknowledgements The authors would like to acknowledge Dr. Jennifer A de Beyer of the Centre for Statistics in Medicine, University of Oxford for editing support.

Compliance with Ethical Standards

Conflict of Interest Sara Khalid declares no potential conflicts of interest. Daniel Prieto-Alhambra reports research grants and advisory and speaker fees paid to his department from Amgen, research grants and

[3] T-score is a measure of bone strength, where a lower T-score indicates worse bone strength than a higher T-score, where a T score = 0 corresponds to the bone mineral density of a healthy adult.

advisory and consultancy fees paid to his department from UCB Biopharma SRL and research grants from Servier, Astellas, Novartis and GSK, outside the submitted work.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as:
•• Of major importance

1. Yadav P, Steinbach M, Kumar V, Simon G. Mining electronic health records (EHRs): a survey. ACM Comput Surv. 2018;50(6):85.
2. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care. 2010;48:S106–S13.
3. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012;13(6):395–405.
4. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc. 2009;16(3):328–37.
5.•• Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform. 2018;22(5):1589–604 **An in-depth review of machine learning methods (with a focus on deep learning) for various stages and aspects of EHR mining, including but not limited to representation learning, phenotyping and prediction**.
6. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2012;20(1):117–21.
7. Chaitanya S, Preethi R, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc. 2014;21(2):221–30.
8.•• Beaulieu-Jones BK, Greene CS. Semi-supervised learning of the electronic health record for phenotype stratification. J Biomed Inform. 2016;64:168–78 **Explains the use of Autoencoder for semi-supervised learning of features of large clinical datasets and describes the validation and visualisation process.**
9.•• Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Reports. 2016;6:26094 **Provides a framework for completely data-driven analysis of EHR data including both phenotyping and prediction stages, using unsupervised and supervised machine learning**.
10. Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PLoS One. 2013;8(6):e66341.
11. NICE. Multimorbidity: clinical assessment and management. NICE guideline, 21 September 2016.
12. Bolíbar B, Fina Avilés F, Morros R, del Mar Garcia-Gil M, Hermosilla E, Ramos R, et al. SIDIAP database: electronic clinical records in primary care as a source of information for epidemiologic research. Med Clin. 2012.
13. Banerjee A, Dave RN. Validating clusters using the Hopkins statistic. IEEE international conference on Fuzzy Systems; 2004.
14. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504–7.
15. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res. 2010;11(Dec):3371–408.
16. Aggarwal CC, Reddy CK. Data clustering: algorithms and applications. 1st ed. Boca Raton: CRC Press; 2013.
17. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability; 1967. Oakland, CA, USA.
18. Johnson SC. Hierarchical clustering schemes. Psychometrika. 1967;32(3):241–54.
19. Bishop C. Pattern recognition and machine learning (information science and statistics). 1st ed. New York: Springer; 2007; 2006. corr. 2nd printing ed.
20. Calinski T, Harabasz J. A dendrite method for cluster analysis. Commun Stat. 1974;3(1):1–27.
21. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc. 2002.
22. Khalid S, Judge A, Pinedo-Villanueva R. An unsupervised learning model for pattern recognition in routinely collected healthcare data; 2018.
23. Pagès-Castellà A, Carbonell-Abella C, Avilés FF, Alzamora M, Baena-Díez JM, Laguna DM, et al. Burden of osteoporotic fractures in primary health care in Catalonia (Spain): a population-based study. BMC Musculoskelet Disord. 2012;13(1):79.