Abstract for ISPE Annual Conference 2018 (Deadline: 14 February 2017)
Current Character Count = 2055/2258, which includes characters in abstract title and abstract body.

Title: Deep Learning for Drug Utilisation Research: Identifying Features Within a Population of Anti-Osteoporosis Drug Users

Authors: Sara Khalid, Klara Berencsi, M Sanni Ali, Peter Rijnbeek, Daniel Prieto-Alhambra

Background
We have previously demonstrated that data-driven cluster analysis can be used to identify subgroups within a population of users of specific drugs, based on key baseline characteristics. Here we demonstrate how deep learning methods can be applied to identify such key features, and to detect sub-groups of interest.

Objectives
To test a data-driven method to learn the features of a population of anti-osteoporosis drug users, and whether these features can be used to derive patient sub-groups.

Methods
Using the SIDIAP Database (anonymized primary care records for >80% of the population of Catalonia), 37,996 incident users of anti-osteoporosis drugs (2007-2014) with complete data on risk factors were analysed.
All available risk factors were included.
A machine learning method called "autoencoder" [1] was used to learn the key features of the population, distributed in *n* clusters. Our previous work using traditional unsupervised learning methods (hierarchical and *k*-means clustering) showed the presence of at least $n = 5$ clusters in this dataset, which we used as reference. The "importance" of a given feature would be determined by its weight P, where the most important feature would have $P = 1$, and unimportant features would have $P = 0$.
Our proposed approach thus provides (a) the important features in the detected clusters, and (c) additionally, the probability of membership of any given patient in each cluster.

Results
At $n = 5$ (the number of plausible sub-groups detected by hierarchical and *k*-means clustering), the autoencoder model detected 9 features to be important ($P \approx 1$): age, gender, body mass index, smoking, alcohol drinking, fracture history, co-morbidity, sedative use, and steroid use. Anticoagulants, aromatase inhibitors, and HRT/contraceptives were not considered important features ($P \approx 0$).

Conclusion
Our study proposed a novel approach for determining the characteristics of a population of AOD users, which is increasingly useful in drug utilisation research where there can be a large number of overlapping risk factors. The features selected by the model were the same as those independently deemed by clinical expert opinion to be important factors considered by GPs when prescribing AODs, which adds face validity to our findings. Further research is needed on the use of this method for drug utilisation research, especially for high-volume, high dimensional data.

References
[1] Cluster Analysis to Detect Drug Utilisation Patterns using Electronic Heath Records. Submitted to 40th International Conference of the IEEE Engineering and Medicine in Biology Society, for publication in IEEE Xplore.

[2] An Unsupervised Learning Model for Pattern Recognition in Routinely Collected Healthcare Data. In Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - Volume 5: HEALTHINF, pages 266-273. SCITEPRESS (2018).

Figures for co-authors' reference (no Figures or Tables allowed to be submitted in Abstract).

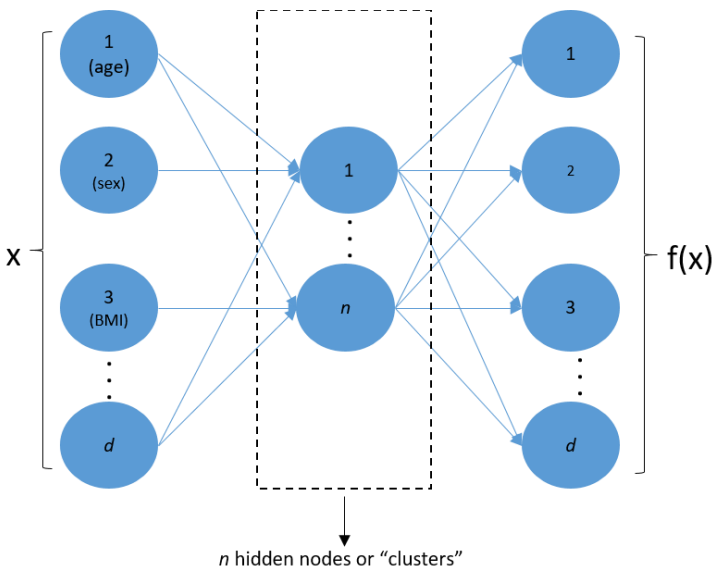


Figure 1. Autoencoder architecture structure.

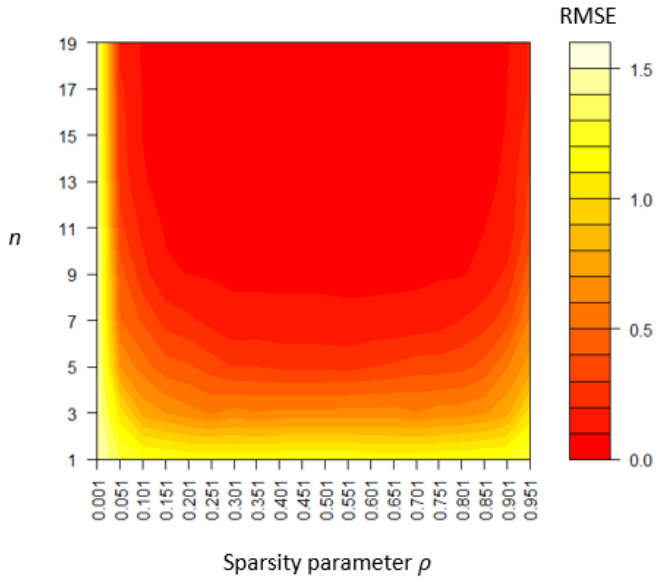


Figure 2. Grid search for optimal number of hidden nodes (n). For $n \geq 5$, RMSE$\leq$ 0.2, indicating that n=5 can appropriately characterise the dataset.

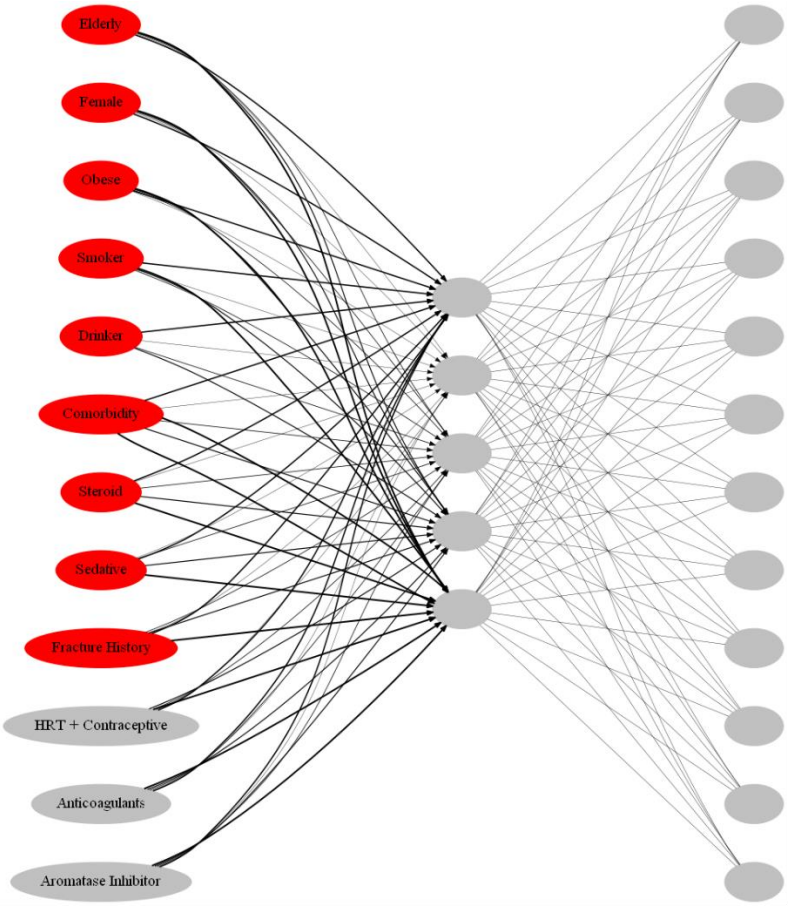


Figure 3. Architecture of optimal autoencoder after learning the features that characterise the dataset. Features detected as "important" are shown in red, while those not "important" are in grey.

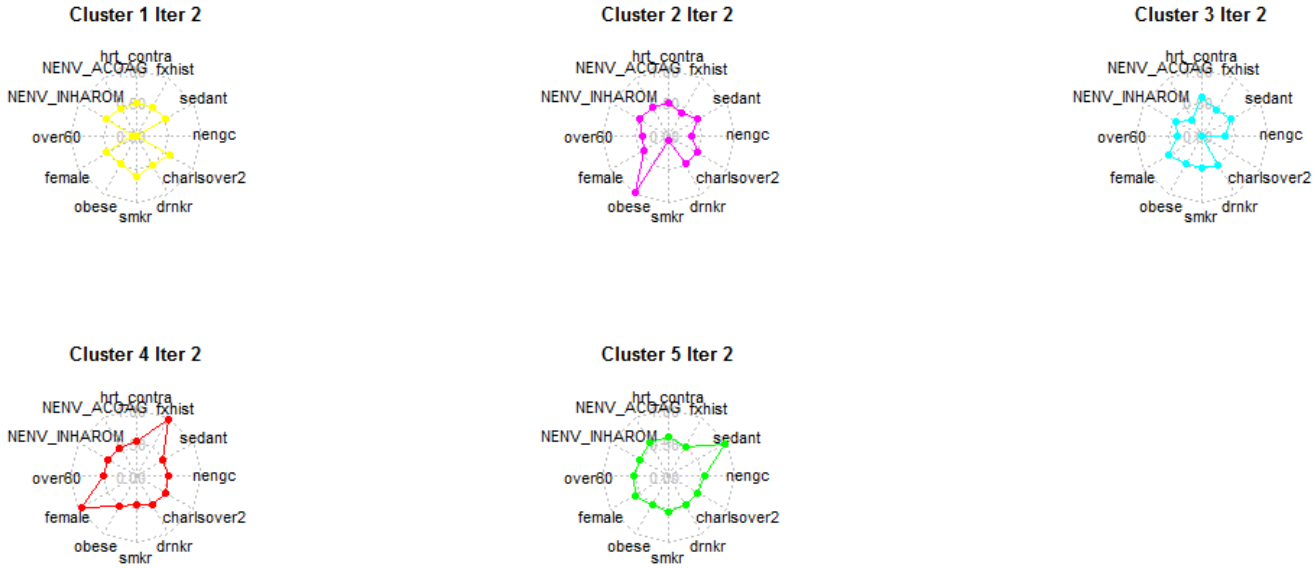
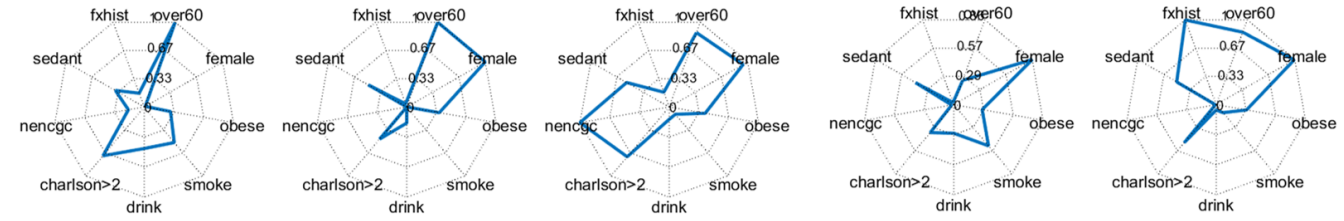
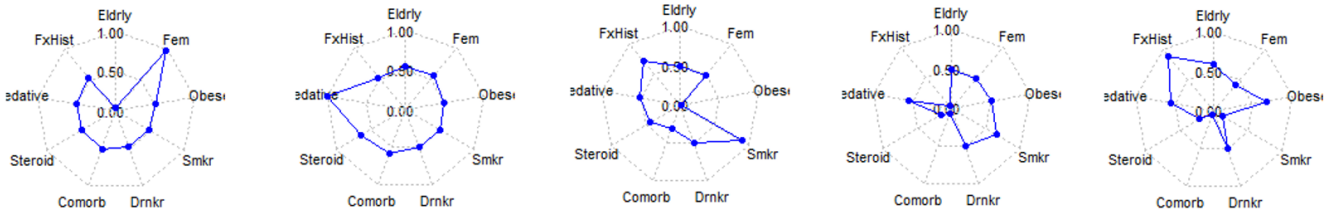


Figure 4. Probability (P) of a feature (or risk factor) at a hidden node (or cluster), where one sub-plot represents one cluster. An important feature would either have P = 1 (highly prevalent), or P =0 (absent). A "non-important" risk factor would take probability values around P=5.



(a) Hierarchical clustering – 5 clusters



(b) Autoencoder feature detection – 5 hidden nodes

Figure 5. Description of 5 clusters obtained using only the 9 features detected by the optimised autoencoder. Clusters obtained using hierarchical clustering also shown (top) as reference.

SK Code List
Sidiap_clusters_fx_risk_2017_01_03.R
AutoencoderScript_SIDIAP.R
Sidiap_cluster_bmd_2016_12_05.m

1. Hinton, G.E. and R.R. Salakhutdinov, *Reducing the dimensionality of data with neural networks.* science, 2006. **313**(5786): p. 504-507.