TITLE PAGE

Title: Machine learning: a long way from implementation in cardiovascular disease

Authors: Suliang Chen[1], Amitava Banerjee[1]

[1]Institute of Health Informatics, University College London

Corresponding author: Dr Amitava Banerjee, Institute of Health Informatics, University College London
222 Euston Road, London,
United Kingdom
NW1 2DA

Corresponding author's email address: ami.banerjee@ucl.ac.uk

Contributorship Statement: The manuscript was conceived by AB. The first draft was jointly prepared by AB and SC. Both authors contributed to revision of the manuscript and have accepted the final version.

Transparency Statement: The corresponding author affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; no important aspects of the study have been omitted.

Data Sharing Statement: No data is available for this study due to ethical constraints.

Funding Statement: AB and SC have received funding from the BigData@Heart Consortium, under the Innovative Medicines Initiative-2(116074, supported by the European Union's Horizon 2020 programme and EFPIA(Chairs: DE Grobbee, SD Anker).

Ethical approval: No ethical approval was not required.

Machine learning: a long way from implementation in cardiovascular disease

The term, “machine learning” (ML), dates back to the 1950s to describe how algorithms and neural network models can assist computer systems in progressively improving their performance. In the last decade, advanced ML algorithms have been increasingly used for phenotypic identification in different cardiovascular diseases (CVDs), driven by two major factors. First, a gap persists between disease definitions from research or consensus guidelines and routine clinical practice. Second, as electronic health records (EHR) are increasingly adopted within and across countries, there are unprecedented opportunities to investigate disease definitions in a more reproducible and generalizable manner. The objective of ML algorithms in such analyses is to develop replicable EHR-based phenotypical definitions for a given CVD(1), and to predict group assignments of new patients.

EHRs include diagnostic codes from primary care, secondary care and administrative data. To produce homogenous subgroups of EHR data, clustering methods are unsupervised ML algorithms which aim to group objects with similar attributes relying on similarity and distance measures. In this issue, Hedman and colleagues(2) identify six phenotypes of heart failure with preserved ejection fraction (HFpEF), derived using data from 320 HFpEF out-patients in the Karolinska-Rennes cohort study. HFpEF is the example par excellence of a disease where disease definitions, and consequently knowledge regarding pathophysiology and management, remain elusive. The team provides a prediction tool for patient assignment of identified subtypes, although this has not been externally validated.

The authors provide new insights in a disease lacking a single evidence-based therapy to alter prognosis. They found that two pheno-groups had worst prognosis: those with hypertension and highest prevalence of coronary artery disease, renal disease, anaemia and diabetes (pheno-group 1) and those with atrial fibrillation and high prevalence of COPD, old age, kidney dysfunction and anaemia (pheno-group 2). Pheno-group 1 is consistent with underlying myocardial dysfunction and what the authors term, “forward failure”, resembling heart failure with reduced ejection fraction, whereas pheno-group 2 showed increased evidence of diastolic dysfunction and right ventricular dysfunction, more in line with HFpEF and a “backward failure” phenotype(2). These phenotypes are consistent with previous studies(3). However, this work also highlights how far we are from routine ML implementation for phenotype discovery, despite much coverage in lay and scientific press suggesting the contrary. Figure 1 summarises the different steps required for ML to reach patient care.

<u>Data</u>

ML is dependent on underlying data for training and validation, which current literature underplays. To-date, published studies investigating CVD phenotypes have been heterogeneous in their study design, including national registry(4), prospective/retrospective cohorts (5), cross-sectional(6) and randomised clinical trials(7). As with traditional epidemiologic studies, design and size matter. A small sample size leads us to question the internal and external validity of the observations and whether the pheno-groups identified are representative of HFpEF patients on the whole. Greater transparency of diagnostic codes and inclusion criteria will lead to greater reproducibility of analyses and results for phenotypic models. In addition, the

way in which missing data and outliers are handled is crucial and should be reported(8) as in this analysis by Hedman and colleagues(2), again facilitating reproducibility of these data. The process by which variables are selected and reduced to a set of principal variables (dimensionality reduction) should be explicitly stated.

Covariates

The number, range and representativeness of covariates is crucial for identification of meaningful phenotypes. Current covariates extracted from EHR include baseline characteristics, comorbidities, laboratory investigations, medications, symptoms and outcomes. Hedman and colleagues add proteomic data to the range of variables linked with particular HFpEF phenotypes. However, variables frequently overlooked by published analyses to-date (including this one) are those unrelated to CVD. Non-CVD variables are often available in patients' historical EHR as part of their longitudinal trajectories, potentially adding to phenotype discovery(9).

The particular processes and criteria for variable selection also need consideration. A small number of patients with a large number of variables can lead to the "curse" of dimensionality and overfitting. Variance within a dataset is also a key factor, and can be mitigated by dimensionality reduction techniques such as principal component analysis for continuous variables and multiple correspondence analysis for categorical variables to remove "noisy" features. There are multiple approaches to reduce dimensionality, but investigators should take into account completeness of available covariates with consideration of size of the patient cohort and whether selected covariates can be linked to study outcomes.

Machine learning

Clustering is important in grouping data into subtypes which are affected by different clustering algorithms and number of clusters. There are many different clustering algorithms, but "hierarchical" (e.g. agglomerative clustering(10)) and "partitioning" (e.g. k-means(3))approaches are most commonly used. Different clustering algorithms have different strengths and weaknesses, necessitating comparative studies to inform algorithm selection before using the method for phenotype discovery(11), but again this aspect is often neglected in published studies. Often just one clustering algorithm which has been applied to the data in each study(3, 10), which may increase the risk of missing important patterns of diseases. Moreover, indices for finding optimal number of clusters should be carefully examined, as different indicators may yield contradictory results, leading to different clusters.

Computing capacity is a key determinant of which ML analyses can and are performed and has been largely neglected by published literature. Ideally, a large cohort with a large number of covariates will increase research generalisability of a given study, but will require greater computing power. On the other hand, the clinical generalisability of such analyses may be limited since such computing power is currently restricted or absent in the clinical context. Despite the substantial opportunities for research from growing datasets with increasing numbers of variables, there are significant challenges for computation, which may be a factor in determining the size of population, number of covariates and the particular ML models used in published phenotype models. For example, calculations which are based on "sample dissimilarity matrix" require more intensive computation. Even when research and clinical institutions upgrade their hardware, generation of usable results is not guaranteed

because many clustering algorithms are not robust enough in large datasets with high dimensionality. Therefore, the majority of phenotypical studies have been limited to smaller or more homogenous datasets(12).

Validation

Just as in models produced by standard regression techniques, the external validity of results from ML-driven analyses is crucial to make findings interpretable and translatable to clinical care. However, the majority of phenotypical studies for CVD to-date only validated results internally(5-7) and not externally (in an independent, prospective dataset). Datasets are usually divided into training and validation sets of pre-specified proportions, and measures of accuracy (e.g. area-under-the-curve, sensitivity and specificity) are calculated. The reason for only conducting internal validation maybe due to the lack of independent cohort, but as both training and validation sets are from same cohort, positive results suggested from such studies are likely to introduce bias, overestimating potential impact on healthcare. External validation itself can be by several methods. First, baseline characteristics may be compared between the derivation and validation cohorts. Second, the predicted and actual clustering of patients in two cohorts can be compared. Third, clusters can be used to predict outcomes, in order to examine the utility of the phenotypes and the possibility of their clinical implementation. There are currently no specific guideline recommendations for validation of ML results in phenotype definition.

Implementation

Even after an ML algorithm is externally validated, several factors need to be addressed before it is ready for clinical implementation. First, outcome data would help to show that the method adds effectiveness to the clinical pathway, whether in a “before-after” study or trial design. Second, the way in which the algorithm will actually fit into a patient pathway should be considered in different settings to understand how the ML application will change current clinical practice, and to understand its feasibility and acceptability with different stakeholders, including health professionals and patients. Finally, perhaps the most significant obstacles to implementation of ML (assuming that it is proven to effective and validated) are external to ML algorithms related to “digital readiness” of the health system (e.g. EHR and IT capacity) and workforce (e.g. training of health professionals in the use of ML). Without consideration of

The road ahead

Hedman and colleagues have made an important contribution in a patient population, where significant clinical and research questions remain. However, their analyses have to be taken in the context of multi-factorial complexity where ML solutions will have to operate. In order to improve the data-driven characterisation of CVD and make impact on clinical decision-making, ML studies for subtyping and risk prediction need to be larger-scale, across diseases, with standardised reporting and validation. Consensus guidelines for ML in research and clinical practice are urgently required if these tools are going to translate to patient care. External validation in research studies using ML in healthcare will help to understand which clustering and prediction tools are of greatest use to the data; and suitable for clinical implementation. In order for ML to create patient benefit, the investigations need to shift from the frameworks of discovery science to evidence-based healthcare and implementation science.

References

1. Alonso-Betanzos A, Bolon-Canedo V. Big-Data Analysis, Cluster Analysis, and Machine-Learning Approaches. Adv Exp Med Biol. 2018;1065:607-26.
2. Hedman A, Hage, C, Sharma, A., Brosnan, M.J., Buckbinder, L.; Gan, L., Shah, S., Linde, C., Donal, E., Daubert, J., Mälarstig, A., Ziemek, D., Lund, L. . Unsupervised machine learning identifies pheno-groups in heart failure with preserved ejection fraction. Heart. 2019.
3. Ahmad T, Lund LH, Rao P, Ghosh R, Warier P, Vaccaro B, et al. Machine Learning Methods Improve Prognostication, Identify Clinically Distinct Phenotypes, and Detect Heterogeneity in Response to Therapy in a Large Cohort of Heart Failure Patients. Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease. 2018;7(8):e008081.
4. Inohara T, Shrader P, Pieper K, et al. Association of of atrial fibrillation clinical phenotypes with treatment patterns and outcomes: A multicenter registry study. JAMA Cardiology. 2018;3(1):54-63.
5. Huang TY, Moser DK, Hwang SL. Identification, Associated Factors, and Prognosis of Symptom Clusters in Taiwanese Patients With Heart Failure. J Nurs Res. 2018;26(1):60-7.
6. Omar AMS, Narula S, Abdel Rahman MA, Pedrizzetti G, Raslan H, Rifaie O, et al. Precision Phenotyping in Heart Failure and Pattern Clustering of Ultrasound Data for the Assessment of Diastolic Dysfunction. JACC: Cardiovascular Imaging. 2017;10(11):1291-303.
7. Riegel B, Hanlon AL, McKinley S, Moser DK, Meischke H, Doering LV, et al. Differences in mortality in acute coronary syndrome symptom clusters. American Heart Journal. 2010;159(3):392-8.
8. Hormozdiari F, Kang EY, Bilow M, Ben-David E, Vulpe C, McLachlan S, et al. Imputing Phenotypes for Genome-wide Association Studies. Am J Hum Genet. 2016;99(1):89-103.
9. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. PLoS One. 2018;13(8):e0202344.
10. Vellone E, Fida R, Ghezzi V, D'Agostino F, Biagioli V, Paturzo M, et al. Patterns of Self-care in Adults With Heart Failure and Their Associations With Sociodemographic and Clinical Characteristics, Quality of Life, and Hospitalizations: A Cluster Analysis. Journal of Cardiovascular Nursing. 2017;32(2):180-9.
11. Bose E, Radhakrishnan K. Using Unsupervised Machine Learning to Identify Subgroups Among Home Health Patients With Heart Failure Using Telehealth. Comput Inform Nurs. 2018;36(5):242-8.
12. Singh N, Garg, N., Pant, J. . A comprehensive study of challenges and approaches for clustering high dimensional data. Int J Comput Appl. 2014; 92:7–10.