Astuces pour l'analyse critique d'article scientifique

Clustering des big data par intelligence artificielle : comprendre les enjeux d'une révolution en marche !

T. PEZEL
Service de Cardiologie,
CHU Lariboisière, PARIS ;
Unité Inserm-UMR 942, PARIS.

L'intelligence artificielle (IA) bouleverse particulièrement le monde de la cardiologie ces dernières années. En effet, l'IA permet le développement d'algorithmes pour l'analyse d'ECG, d'images d'échocardiographie ou d'IRM cardiovasculaire.
Au-delà de nous permettre d'interpréter ces données techniques, l'IA est également utile pour l'analyse statistique de grandes bases de données comportant plusieurs milliers de patients avec des dizaines de paramètres mesurés : la fameuse "gestion des big data".
Ces toutes dernières années, de nombreuses études cardiologiques ont été publiées à très haut niveau, utilisant ce type d'outil d'IA et permettant de mieux classer les patients en fonction de la meilleure analyse possible des données disponibles.
Le *clustering* est incontestablement l'entité la plus fréquemment rencontrée avec ce type d'outil. Ainsi, dans cet article nous présenterons, de façon simple, les grands principes du *clustering* dans le but de nous permettre d'analyser les résultats d'une étude utilisant ce type de méthode.

Principe du clustering

Le ***clustering*** est une méthode d'analyse statistique utilisée pour **organiser des données brutes sous la forme de groupes les plus homogènes possibles : les *clusters***. À l'intérieur de chaque groupe (*cluster* ou phénogroupe), les données sont regroupées selon certaines caractéristiques communes identifiées par un outil d'IA. En français, on emploie couramment le terme de "regroupement de données".

Cet outil correspond à un algorithme qui mesure la distance entre chaque patient dans un espace construit à partir de certaines variables prédéfinies (âge, sexe, présence d'un diabète, d'un antécédent d'infarctus du myocarde…). Une fois que cet outil a mesuré cette "distance" entre les patients, il tente de les classer au mieux afin de créer autant de *clusters* qu'il semble nécessaire tout en minimisant au maximum cette distance.

Au final, on obtient une figure permettant de visualiser l'appartenance de chaque patient à un *cluster* tout en appréciant à quel point le *clustering* est performant avec des *clusters* séparés les uns des autres (***fig. 1***).

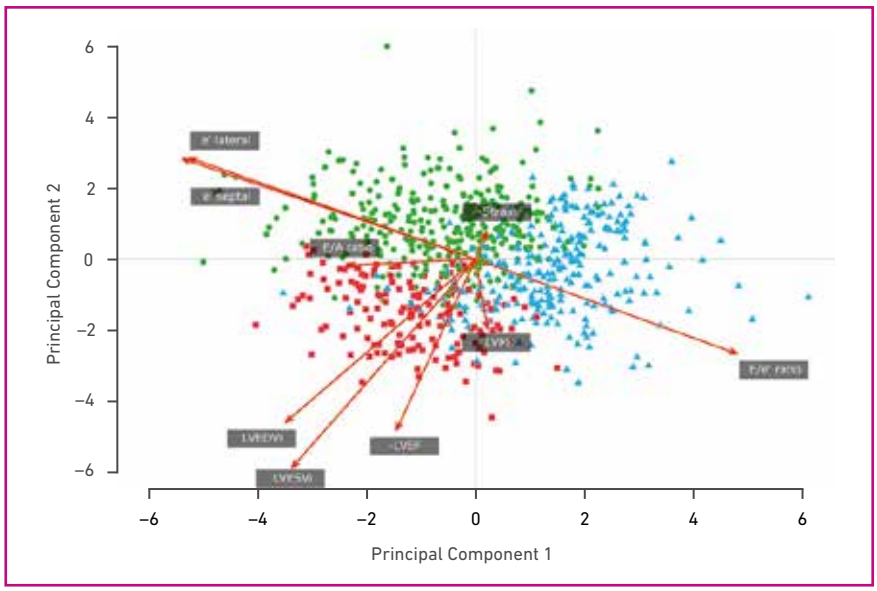

Fig. 1 : Exemple de représentation dans l'espace de chaque patient en fonction d'un ensemble de données prédéfinies mesurées en échocardiographie (d'après Ernande *et al. JACC*, 2017). Chaque patient est représenté par un rond vert s'il appartient au *cluster* 1 (**vert**), par un triangle bleu s'il appartient au *cluster* 2 (**bleu**) ou par un carré rouge s'il appartient au *cluster* 3 (**rouge**). Chaque variable mesurée en échocardiographie est représentée par une vignette portée par un vecteur permettant d'évaluer à quel point une variable participe à la constitution de chaque *cluster*. Ainsi, on peut par exemple dire que la valeur du "ratio E/e'" tout à droite de la figure est une variable majeure pour définir le *cluster* 2 (**bleu**) situé du même côté.

Qu'appelle-t-on une "méthode d'analyse non supervisée" ?

La méthode d'analyse non supervisée la plus fréquemment utilisée est **l'analyse en composantes principales (ACP).** L'ACP est un outil mathématique qui permet de réduire le nombre de variables en trouvant de **nouvelles variables qui en synthétisent plusieurs.** Ainsi, trouver une variable synthétique permet de remplacer plusieurs anciennes variables par une seule et donc **facilite le regroupement de patients (*clustering*)** ensuite. Cependant, cette transformation entraînera forcément une perte d'information.

Différence entre méthode d'analyse non supervisée ou supervisée :

>>> La **classification non supervisée** consiste en l'organisation d'individus en groupes homogènes à partir d'une population hétérogène mal définie. En résumé, on caractérise des classes que l'on ne connaît pas à l'avance avant le début de l'étude.

>>> La **classification supervisée** consiste à "ranger" les individus dans des classes connues.

Pour réaliser cette "simplification" des variables, l'ACP utilise grossièrement deux mesures :

– la variabilité entre les patients, c'est-à-dire quelles sont les différences et les ressemblances entre les patients de la population totale ;
– les liaisons entre les variables : existe-il des groupes de variables très corrélées entre elles qui peuvent être regroupées en de nouvelles variables synthétiques ?

Comment apprécier la qualité d'un clustering lorsqu'on lit l'article ?

L'enjeu d'un *clustering* est de proposer une classification suffisamment performante afin de distinguer, de façon claire, les différents groupes en fonction des variables mesurées.

Ainsi, le premier outil permettant d'évaluer facilement cette distinction entre les *clusters* et les potentielles zones de chevauchement (*overlaps*) est le "***radar-chart***" (***fig. 2***).

Bien que ce type de figure permette une visualisation rapide et facile des caractéristiques de chaque *cluster*, cela ne

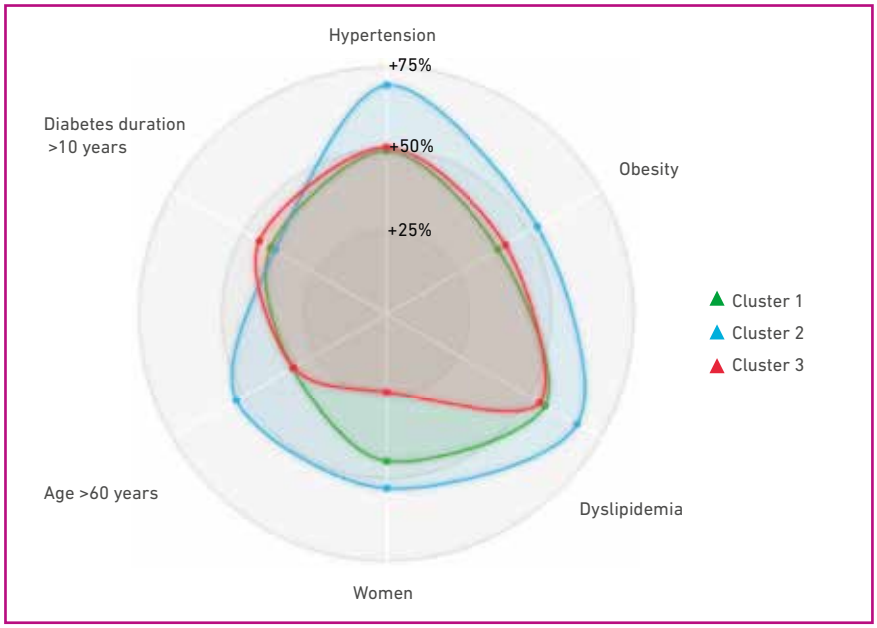

Fig. 2 : Exemple de *radar-chart* présentant la répartition de 3 *clusters* selon les variables cliniques mesurées dans l'étude (d'après [1]). On note ici qu'il existe par exemple un chevauchement complet de l'hypertension entre les *clusters* 1 et 3, alors que le *cluster* 2 présente le taux de patients avec hypertension le plus élevé.

Astuces pour l'analyse critique d'article scientifique

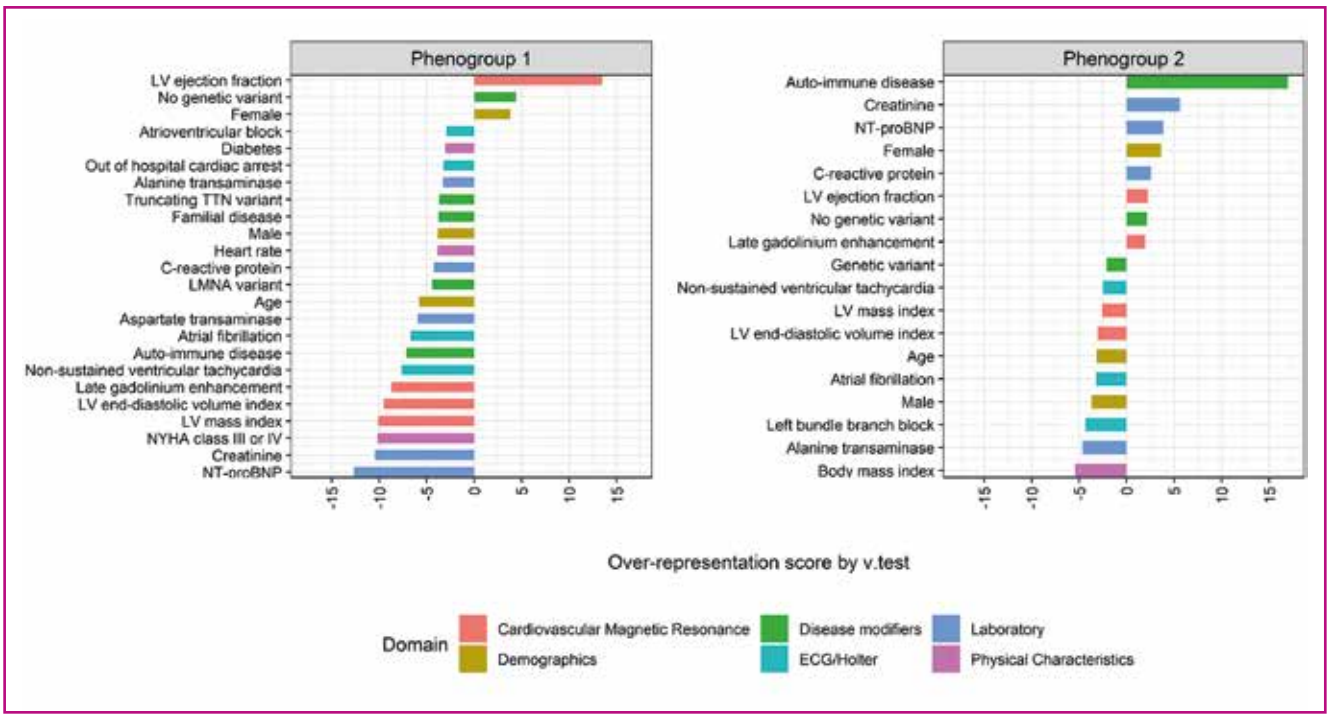

Fig. 3 : Représentation de la participation de chaque variable dans la création des *clusters* en utilisant le "v. test" (d'après [2]). Une valeur positive indique une surreprésentation de la variable dans le *cluster* étudié, alors qu'une valeur négative indique une sous-représentation de la variable correspondante. Quel que soit le signe de la valeur, plus cette valeur est élevée et plus cela signifie que cette variable participe à la constitution du *cluster*. De plus, notez les nombreuses catégories de variables utilisées : variables cliniques, biologiques, génétiques, démographiques, d'ECG/Holter et d'IRM cardiaque.

permet pas une mesure précise de la composante de chaque variable dans la caractérisation de chaque *cluster*. C'est pour cette raison que les auteurs de ce type d'étude utilisent de plus en plus souvent un outil statistique dédié appelé le "v. test". Son interprétation est simple : "plus le v. test d'une variable dans un *cluster* est élevé et plus cela signifie que cette variable est importante dans la caractérisation du *cluster*" (***fig. 3***).

Finalement, dans quel type d'étude un clustering peut-il être utile ?

La méthode statistique utilisant une analyse non supervisée pour réaliser du ***clustering*** **doit être appliquée sur une population hétérogène que l'on a du mal à définir.** Par exemple, Shah *et al.* ont été les premiers à utiliser ce type d'outil pour caractériser l'insuffisance cardiaque à FEVG préservée [3]. C'est un magnifique exemple d'utilisation du *clustering* car l'insuffisance cardiaque à FEVG préservée correspond bien à un ensemble de patients hétérogène difficile à stratifier et encore partiellement mal défini.

Une fois que la population à caractériser est bien définie, il est important d'évaluer le maximum de variables possibles permettant de caractériser ces patients. De plus, il est recommandé d'**étudier des catégories de variables les plus différentes possibles :** variables cliniques, biologiques, génétiques, sociologiques, d'ECG ou de Holter, d'échocardiographie, d'IRM cardiaque (***fig. 3***)…

L'importance d'évaluer la valeur pronostique du clustering

Lorsqu'on réalise un *clustering* comme décrit précédemment, on peut parfaitement s'arrêter à une étape purement descriptive qui correspondrait au fait de décrire une population très hétérogène sous la forme de *clusters* avec une signification clinique éventuelle. On parle d'études "à visée exploratoire".

Cependant, le plus souvent, les auteurs vont plus loin en recherchant des différences de pronostic entres les *clusters* ! En effet, dans ce cas, **les *clusters* présentent des valeurs pronostiques distinctes,** ce qui aide les cliniciens à **mieux stratifier le risque d'événements cardiovasculaires** dans cette population initialement hétérogène et mal définie. Dans cette situation, on utilise presque toujours une courbe de survie de Kaplan-Meier permettant de comparer les courbes entre chaque *cluster* (***fig. 4***).

Enfin, une meilleure stratification du risque permet d'identifier les patients les plus à risque et donc de **guider la prise en charge diagnostique et thérapeutique.**

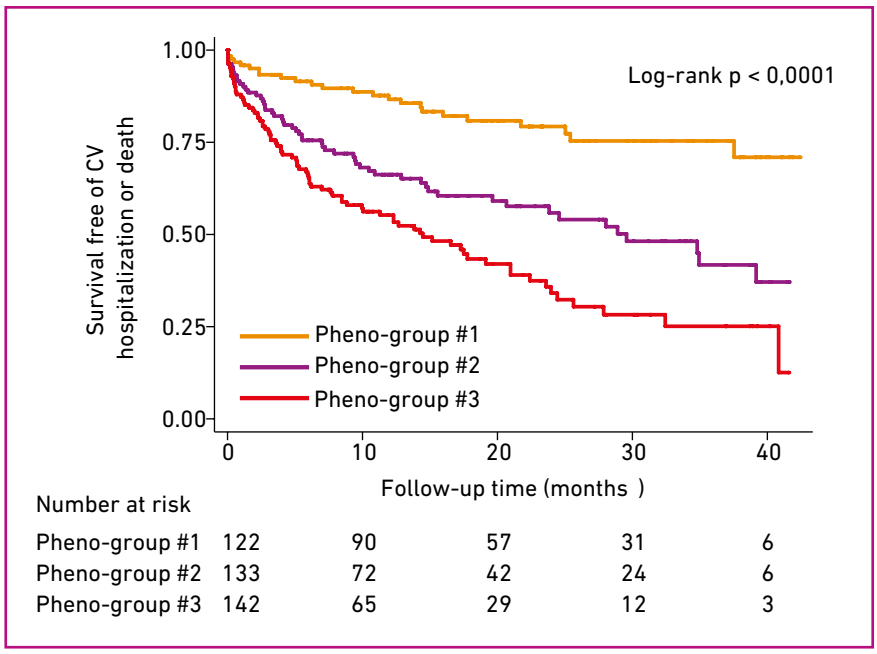

Fig. 4 : Courbe de survie de Kaplan-Meier classique évaluant la survenue d'hospitalisations pour cause cardiovasculaire ou décès dans chaque *cluster* (= phénogroupe). Cette figure permet de comparer la valeur pronostique de chaque *cluster* à l'aide d'un *log-rank test*.

Conclusion

Le *clustering* est un outil d'IA très puissant pour nous permettre de mieux comprendre des populations hétérogènes que l'on a du mal à définir comme, par exemple, les insuffisants cardiaques à FEVG préservée, les patients avec une hypertension pulmonaire ou une cardiomyopathie dilatée.

Son intérêt majeur réside dans le fait de nous permettre de voir des "réalités cliniques associatives" que l'esprit humain seul n'arrive pas à synthétiser. En effet, le développement des techniques est tel à notre époque que le nombre de données obtenues pour chaque patient nécessite l'utilisation de ces analyses non supervisées pour être associées puis percées à jour.

Enfin, l'objectif final de ces travaux est de mieux stratifier le risque de ces populations hétérogènes afin de proposer ensuite des études thérapeutiques randomisées de qualité.

BIBLIOGRAPHIE

1. Ernande L, Audureau E, Jellis CL *et al.* Clinical implications of echocardiographic phenotypes of patients with diabetes mellitus. *J Am Coll Cardiol*, 2017;70:1704-1716.
2. Verdonschot JAJ, Merlo M, Dominguez F *et al.* Phenotypic clustering of dilated cardiomyopathy patients highlights important pathophysiological differences. *Eur Heart J*, 2020. doi:10.1093/eurheartj/ehaa841.
3. Shah SJ, Katz DH, Selvaraj S *et al.* Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation*, 2015;131:269-270.