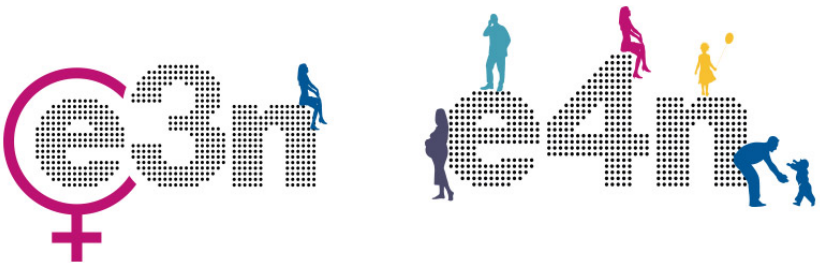
CentraleSupélec

Patterns de multi-morbidité et facteurs promoteurs ou protecteurs dans la cohorte longitudinale de femmes française E3N

10/06/2020

Rapport final



Equipe n° 2:
Léa ABDALLAH
Apolline DERSY
Youssef IRHBOULA
Zoé KANOUNNIKOFF
Camille SCHNEIDER

Institut Gustave Roussy

Contents

1 Introduction

Notre enseignement d'intégration, en collaboration avec l'Institut Gustave Roussy, a pour sujet l'étude des patterns de comorbidité dans la cohorte de femmes françaises E3N. Nous avions à notre disposition une immense base de données regroupant les vécus médicaux d'environ 100 000 femmes, sur une trentaine d'années. Il s'agit de la plus grande cohorte longitudinale française. Du fait de sa taille, cette base de données présente de nombreuses données manquantes qu'il nous a fallu traiter.

Cette base de données se compose de deux groupes : le groupe des facteurs d'exposition (taille, IMC, facteurs hormonaux, alimentation, tabac, niveau d'études...) et le groupe des maladies. Une première analyse de la base de données nous a permis de remarquer une corrélation entre certaines de ces maladies; l'objectif de notre étude a ensuite été de déterminer un lien entre facteurs d'exposition et présence de co-morbidités (i.e. de plusieurs maladies à la fois) chez les femmes de la cohorte E3N.

2 Gestion des données manquantes

2.1 Cas des variables indicatrices

Nous avons considéré pour ces variables que les données manquantes correspondaient à des personnes non concernées. On les remplace alors par des 0. Cela a concerné les variables de maladies excepté l'asthme et la migraine dont nous discuterons le cas ultérieurement.

2.2 Cas des variables catégorielles

Nous avons opté pour un tableau disjonctif complet pour pallier la difficulté inhérente aux variables catégorielles. Les données manquantes ont été remplacées par les médianes.

Toutefois, concernant l'asthme et la migraine, il nous a semblé pertinent de réunir les niveaux 1 et 2 (respectivement 2 et 3) au sein d'une seule variable indicatrice. En effet, migraine1 et migraine2 (respectivement asthme2 et asthme3) ont des corrélations très similaires avec les autres variables; les réunir permet alors de mieux constater leur significativité.

2.3 Cas des variables quantitatives

On remplace les données par la moyenne globale dans de nombreux cas, ou par un calcul plus pertinent le cas échéant. On remplace par exemple la durée d'allaitement par la durée moyenne d'allaitement par enfant, multipliée par le nombre d'enfants. Le cas de la ménopause était également particulier puisque nous avons dû distinguer le cas des femmes non-ménopausées, en cours de ménopause et ménopausées pour leur attribuer un âge de ménopause.

2.4 Exemple de la colonne MENO

MENO peut prendre 4 valeurs:

1. La femme est ménopausée
2. La femme n'est pas encore ménopausée

3. La femme est en cours de ménopause

4. La femme n'a jamais eu ses règles

Données manquantes
Nous avons calculé quelques pourcentages pour nous permettre de compléter les cases vides par des données réalistes. Nous avons trouvé que 88% des femmes dont ageq1 $<$ 48.11 ne sont pas encore ménopausées (MENO = 2), 89% des femmes dont ageq1$>$53 sont ménopausées (MENO = 1), et les femmes dont 48.11 $\leq$ ageq1 $\leq$ 53 sont réparties également entre MENO = 1, MENO = 2 et MENO = 3.

- Si la colonne MENOAGE est également vide, on remplit MENO en fonction de ageq1 (1 si ageq1 $<$ 48.11, 2 si ageq1$>$53, et 3 sinon).
- Si un âge de ménopause est indiqué, on choisit MENO = 1.

Données incohérentes
Certaines données sont parfois incohérentes. Il arrive ainsi de trouver des femmes non ménopausées (MENO = 2) ou n'ayant jamais eu leurs règles (MENO = 4) pour lesquelles un âge de ménopause est indiqué. Dans ce cas, si ageq1-MENOAGE $>$ 7, on peut considérer que la femme est ménopausée (MENO = 1) ; si ageq1-MENOAGE $\leq$ 7 on considèrera que la femme est en cours de ménopause (MENO = 3).

3 Description théoriques des modèles

Lors de notre étude, nous avons utilisé plusieurs méthodes statistiques, notamment l'**analyse en composantes principales** et le **clustering par ascendance hiérarchique** .

3.1 Analyse en composantes principales

Notre jeu de données comportant beaucoup de variables, il nous a semblé très utile d'effectuer une analyse en composantes principales (ACP). En effet, l'ACP est une méthode de projection de données d'un espace de grande dimension p vers un espace de plus petite dimension k. L'objectif est de réduire la dimension des données tout en gardant un maximum d'information, qui correspond ici à la variance de l'échantillon. Dans le cas où les deux première composantes principales expliquent une part suffisante de la variance totale, on peut représenter le jeu de données dans un espace de dimension 2; l'interprétation est alors très facilitée.

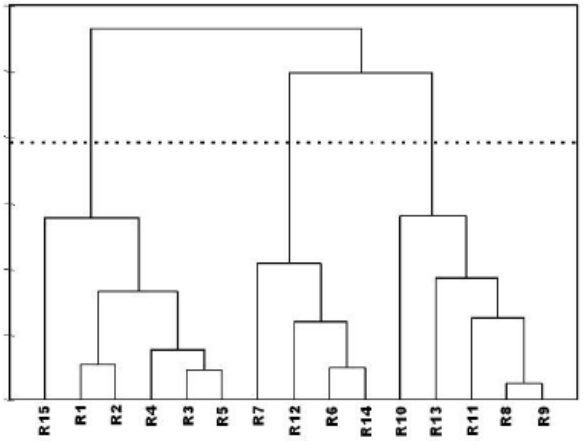
3.2 Clustering par ascendance hiérarchique

Le clustering par ascendance hiérarchique (CAH) est une méthode de classification de données qui renvoie un **dendrogramme**. L'objectif de cette méthode est de détecter des groupes d'individus parmi les données.
La mise en place de l'algorithme nécessite la définition d'une distance entre individus (distance euclidienne, matrice de similarité...) et entre groupes d'individus. On peut par exemple choisir la distance entre les deux individus les plus éloignés du groupe, ou la distance de Ward, qui minimise à chaque étape la perte d'information.
Le principe de l'algorithme est le suivant :

- On cherche les deux points les plus proches du jeu de données (selon la distance choisie) et on les regroupe dans un cluster. Les deux points sont alors remplacés par leur centre.
- A chaque étape, on cherche les deux points (ou clusters) les plus proches et on les regroupe en un cluster.
- On procède ainsi jusqu'à ce qu'il n'y ait plus qu'un seul cluster.

L'algorithme renvoie un dendrogramme, comme celui de la Figure 1. Pour former des clusters, il faut alors couper le dendrogramme à une certaine hauteur; les points étant sur une même branche du dendrogramme forment alors un cluster.

Figure 1: Un exemple de dendrogramme



Cette méthode a l'avantage d'être déterministe, contrairement à la méthode des K-means. Cependant, elle dépend beaucoup de la distance choisie et le choix de la hauteur à laquelle couper n'est pas toujours évident.

3.3 Clustering par K-means

K-means est un algorithme de clustering qui repose sur le **partitionnement** des données. Comme pour le CAH, l'algorithme nécessite la définition d'une distance entre individus et entre groupes d'individus.
A la différence du CAH, K-means prend en entrée le nombre K de clusters à créer en plus du jeu de données. L'algorithme fonctionne alors de la façon suivante :

- A la première étape, on choisit aléatoirement K points qui sont les premiers centres des clusters
- A chaque étape, on associe chaque point au cluster dont il est le plus proche du centre
- On recalcule le centre de chaque cluster
- On recommence les deux dernière étapes jusqu'à ce qu'il y ait convergence, ou stabilisation de l'inertie totale du jeu de données.

A noter qu'on peut également fixer le nombre d'itérations de l'algorithme, au lieu de le faire tourner jusqu'à ce qu'il converge.
L'inconvénient majeur de cet algorithme est son non-déterminisme, dû au choix aléatoire de points à la première étape. On peut ainsi tomber sur un minimum local de l'inertie, et donc ne pas obtenir un partitionnement optimal des données.

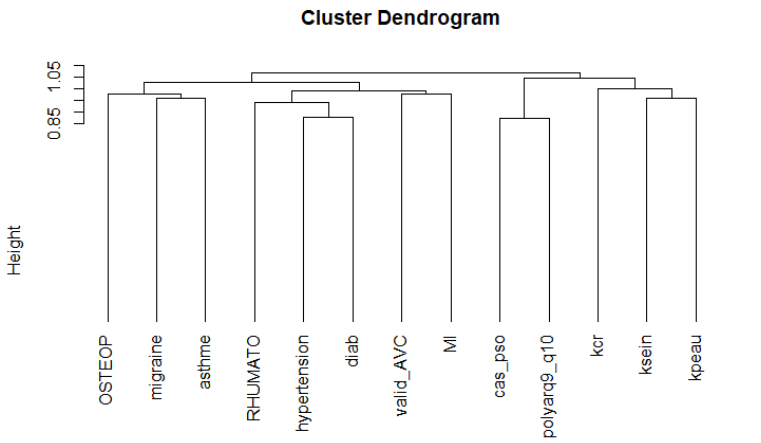
4 Présentation des résultats, figures, commentaires des résultats

4.1 Relation entre les maladies

4.1.1 Clustering hiérarchique

Nous avons réalisé un clustering par ascendance hiérarchique sur les maladies, à l'aide de la librairie R ClustOfVar. [1] A l'aide du code R disponible en annexe, on obtient le dendrogramme de la Figure 2.

Figure 2: Dendrogramme des maladies.



On peut alors répartir les maladies en 5 clusters :

1. Psoriaris et polyarthrite
2. AVC, hypertension, infarctus, rhumatisme et diabète
3. Migraine, asthme et osthéoporose
4. Cancer du sein et cancer de la peau
5. Cancer colorectal.

Cette première représentation nous a été très utile, puisqu'elle nous a permis d'identifier des groupes de maladies qui revenaient souvent ensemble, l'exemple le plus frappant que nous ayons remarqué étant le duo psoriasis/polyarthrite. Par la suite, nous avons étudié les facteurs influençant la probabilité d'appartenir à un de ces 5 clusters.

[1] `https://www.rdocumentation.org/packages/ClustOfVar/versions/1.1`

4.1.2 Visualisation sous forme de graphe

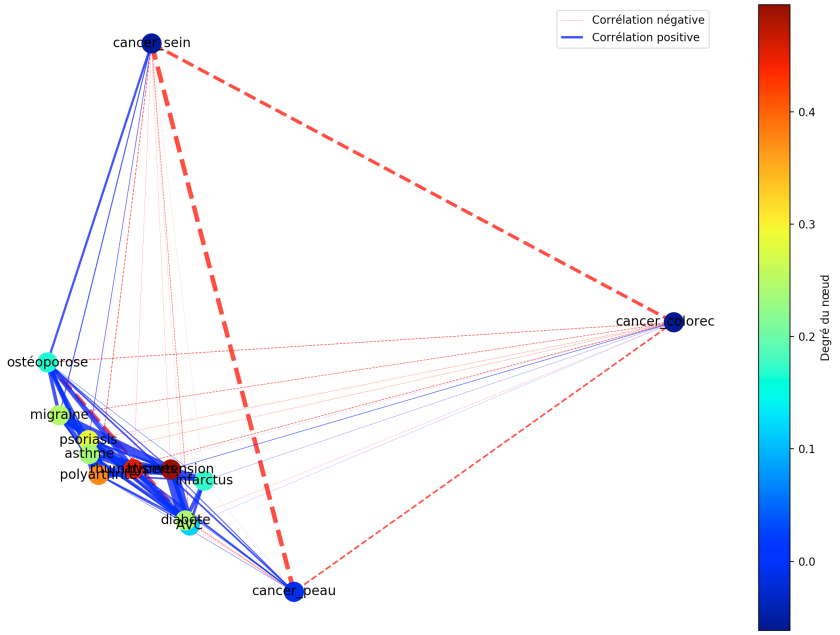


Figure 3: Représentation des maladies sous forme de graphe

Nous avons utilisé le package NetworkX[2] afin de visualiser les relations entre les événements de santé sous la forme d'un graphe représenté Figure 3. Chaque maladie est représentée par un nœud. Les arêtes relient deux événements de santé entre eux avec un poids correspondant à la corrélation entre ces maladies. La matrice de corrélation des événements de santé a été calculée avec la fonction cor de R[3], et est représentée Figure 4.

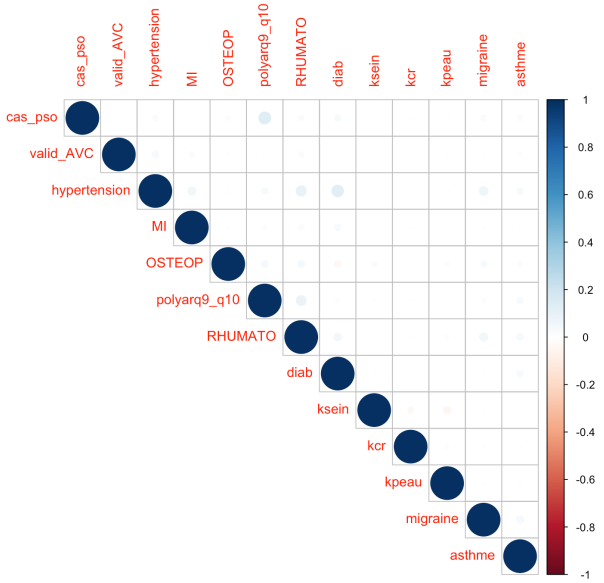


Figure 4: Représentation de la matrice de corrélation des événements de santé

[2] `https://networkx.github.io/documentation/stable/index.html`

[3] La fonction cor de R calcule les coefficients de corrélation en utilisant la méthode de Pearson ou de Spearman. La norme de la différence de ces deux matrices de corrélation étant de l'ordre de 10^{-31}, on peut considérer que ces deux méthodes sont ici équivalentes. Il est important de noter que les deux méthodes s'intéressent uniquement aux corrélations linéaires entre variables.

Le poids de l'arête, et donc la corrélation entre deux événements de santé, est symbolisé par l'épaisseur, la couleur et le style du trait : une corrélation positive est tracée en trait plein bleu, tandis qu'une corrélation négative est tracée en pointillés rouges. Plus le trait est épais, plus la valeur absolue de la corrélation est grande.

La position des points dans l'espace est d'abord choisie aléatoirement, puis NetworkX utilise l'algorithme de Fruchterman-Reingold [1] pour placer les points. Les arcs sont assimilés à des ressorts dont la constante de raideur dépend du poids de l'arc, et les nœuds à des charges positives qui se repoussent. L'algorithme cherche à minimiser l'énergie du système. Ainsi on obtient une représentation visuelle dans laquelle les points les plus éloignés sont les moins corrélés.

On remarque que les cancers sont très peu, voire négativement, corrélés avec le reste des maladies. Visuellement, on peut retrouver les 5 clusters décrits page 4.

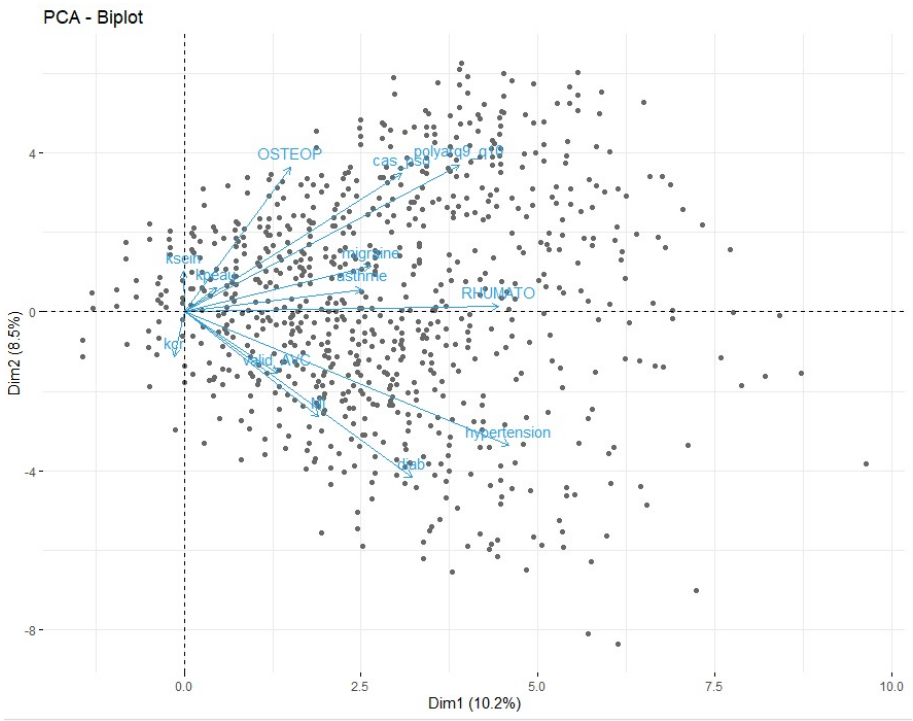
4.2 Clustering des patients

Pour coder cette partie, nous avons utilisé les aides du site [2].

4.2.1 Analyse en composantes principales

Nous avons fait une analyse en composantes principales en prenant comme variables uniquement les maladies. Pour ce faire on utilise la librairie factoextra et le code qui est en annexe. La fonction PCA centre réduit les données elle-même, il n'y a donc pas besoin de le faire avant d'exécuter cette fonction.

Figure 5: Biplot de l'ACP



On voit sur le biplot les points qui représentent chaque individu et les vecteurs des variables.

4.2.2 Valeurs propres de l'ACP

On peut ensuite regarder quel est le pourcentage expliqué pour chaque valeur propre :

numéro valeur propre	valeur propre	pourcentage de variance expliqué	pourcentage de variance expliqué cumulé
1	1.32	10.16	10.16
2	1.11	8.53	18.69
3	1.05	8.11	26.80
4	1.04	8.03	34.83
5	1.01	7.79	42.61
6	1.00	7.75	50.36
7	0.98	7.54	57.90
8	0.97	7.46	65.37
9	0.95	7.32	72.69
10	0.94	7.24	79.94
11	0.92	7.06	87.00
12	0.85	6.56	93.56
13	0.84	6.43	100.00

Une valeur propre > 1 indique que la composante principale concernée représente plus de variance par rapport à une seule variable d'origine, lorsque les données sont standardisées (ce qui est le cas ici). Ici, les six premières valeus propres sont supérieures à 1. On pourrait donc conserver ses six premières valeurs propres pour expliquer nos données. Seulement, dans l'idéal, les valeurs propres choisies doivent avoir un pourcentage de variance expliqué en cumulé supérieur à 75%. On voit bien que ce n'est pas le cas ici et qu'il faudrait prendre 10 valeurs propres pour que ça soit le cas, ce qui revient presque à considérer le problème de départ à 13 dimensions.
On va donc considérer pour la suite les six premières dimensions.

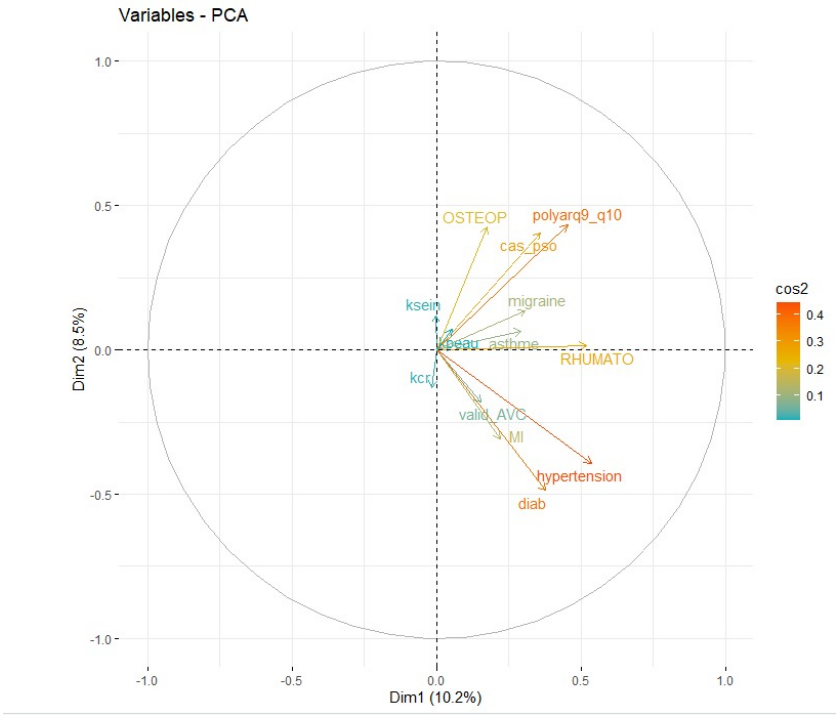
L'ACP ici n'est donc pas très pertinente puisque aucune valeur propre n'explique assez la variance. Mais on peut quand même utiliser cet ACP pour dégager des tendances dans la population.

4.2.3 Qualité des représentations de l'ACP

On peut ensuite regarder la qualité des représentations de l'ACP avec le cos2.
La distance entre les variables et l'origine mesure la qualité de représentation des variables. Les variables qui sont loin de l'origine sont bien représentées par l'ACP. En effet, un cos2 élevé indique une bonne représentation de la variable sur les axes principaux en considération. Dans ce cas, la variable est positionnée à proximité de la circonférence du cercle de corrélation.Un faible cos2 indique que la variable n'est pas parfaitement représentée par les axes principaux. Dans ce cas, la variable est proche du centre du cercle.

On voit donc que les maladies comme l'hypertension, le diabète, la polyartrite sont très bien représentées. Cela semble logique puisque ce sont les maladies que les femmes ont le plus dans nos données, on a donc beaucoup de données sur ces maladies. Au contraire, les trois cancers sont les moins bien représentés car ce sont ceux avec le moins de cas de malades.

Figure 6: cos2 de chaque variable



4.2.4 Poids de chaque variable sur les axes de l'ACP

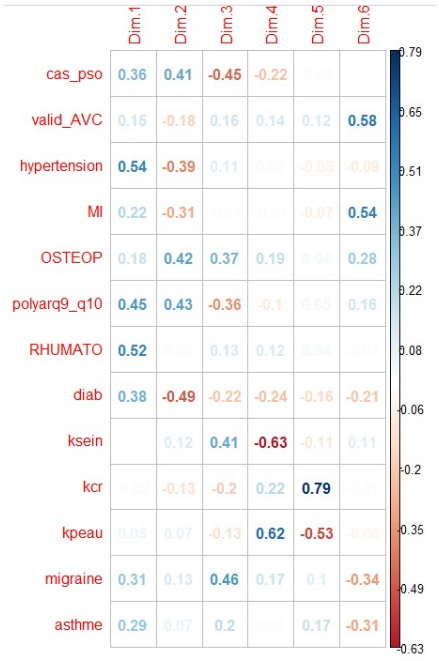


Figure 7: Coordonnées de chaque variable sur les axes de l'ACP

On peut aussi regarder le poids de chaque variable sur les axes de l'ACP. La dimension 1 permet d'associer les maladies aux plus hautes valeurs à un mauvais état de santé et à une tendance à la comorbidité. En effet: rhumato hypertension et migraines ont été identifiés comme les maladies les plus courantes dans toutes les combinaisons de comorbidité et sont souvent elle mêmes associées. De même le psoriasis est souvent présent en comorbidité: dans les trio de maladies rares identifiés comme probables, 12/20 comportent le psoriasis.

La dimension 2 sépare des clusters de maladies: on remarque que psoriasis, polyarthrite et ostéoporose sont proches, et s'opposent au cluster hypertension diabète MI.

On peut retrouver en annexe 20 les valeurs des contributions de chaque variable sur chaque dimension mais cela est moins pertinant que les poids.

4.2.5 Clustering par k-means à partir des coordonnées de l'ACP

A partir de cette ACP, on va pouvoir établir des clusters de personnes.
On va réaliser la méthode du k-means pour créer des clusters de population. Pour constituer nos clusters, on va s'appuyer sur les coordonnées des points des données dans notre ACP.
On décide de prendre 5 clusters de personnes comme ceux réalisés dans la partie 5.1. Avec 5 clusters, le pourcentage d'inertie expliqué est de 48,6%. Ce n'est pas une valeur très élevée, mais nos données ici servent à dégager des tendances et non a expliquer totalement des maladies (ce qui n'est pas possible réellement). D'après la figure 21 données en annexe, on pourrait prendre plus de clusters pour augmenter le pourcentage d'inertie. Mais cela n'aurait pas grand intéret puisque l'on verrait des clusters se ressemblant énormément comme nous vous l'avions montré lors d'une présentation. On voit cela sur la figure 22 avec 9 clusters.
On obtient le biplot suivant :

Figure 8: Biplot des clusters sur l'ACP

On voit que l'on a 3 gros clusters plutot distincts, et 2 clusters inclus dans les autres. On pourra donc bien utiliser les résultats obtenus pour les clusters 1, 2 et 4 et prendre avec moins de considération les résultats des clusters 3 et 5. Si on refait tourner le k-means avec le même nombre de cluster, on a presque toujours les mêmes clusters qui se dégagent.

On va ensuite regarder la proportion d'individus atteints par chaque maladie dans chaque cluster. On constate alors les caractérisations suivantes:

On voit donc :
Cluster 1: 100% de cancer colorectal mais en dessous de la moyenne pour les autres maladies
Cluster 2: psoriasis très au dessus de la moyenne, souvent accompagné de polyarthrite, absence d'infarctus, AVC, cancer colorectal, diabète

Figure 9: Pourcentage de présence d'une maladie par cluster

	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
cas_pso	8.02583026	86.05769231	0.000000	9.0684254	0.2729357
valid_AVC	0.73800738	0.00000000	0.000000	49.7938994	0.0000000
hypertension	52.49077491	54.53725962	32.514561	79.8021434	85.8109829
MI	0.00000000	0.00000000	0.000000	51.7724650	0.0000000
OSTEOP	9.68634686	14.48317308	10.829746	13.6850783	12.4677026
polyarq9_q10	1.29151292	21.15384615	0.000000	4.4517725	0.0000000
RHUMATO	25.27675277	31.15985577	9.752680	37.4278648	55.0747844
diab	6.18081181	0.09008805	1.429898	13.1079967	17.0566615
ksein	0.05260015	8.49358974	8.366675	7.3371805	8.5155937
kcr	100.00000000	0.00000000	0.000000	0.7419621	0.0000000
kpeau	0.00000000	2.20352564	1.814805	1.8961253	2.3326904
migraine	22.23247232	28.11498397	10.836499	26.2984336	53.4699225
asthme	0.69783908	0.93877141	2.888495	11.4591921	19.1710033

Cluster 3: groupe sain, nombreuses maladies absentes, maladies courantes moins présentes que la moyenne
Cluster 4: cluster AVC et infarctus, plus haut en hypertension rhumato asthme et diabète
Cluster 5: cluster rhumato migraine diabète asthme hypertension en absence d'AVC MI psoriasis et polyarthrite

Puis on va regarder les valeurs moyennes des facteurs d'exposition centrés réduits pour chaque cluster.

Figure 10: Ecart relatif à la moyenne de toutes les femmes de l'étude par facteur et pour chaque cluster

	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
imcq1	1.0870182	1.137904e+00	-2.0883229	3.8480417	3.8755039
ageq1	5.0931574	-1.509672e-01	-1.6732953	8.8507609	3.0702700
soleil	-3.6606202	1.037798e+00	-0.7534005	-1.7389239	1.4681657
encoupleQ1	0.8656955	-1.411960e+00	0.9631802	-6.4676969	-1.3119133
age1ergross	2.6920852	2.536916e+00	2.8409571	1.6184307	2.0824700
nbenfvivmn	1.6298057	-1.432662e+00	-0.9737784	5.8971887	2.2950418
allaitement_dureecum	1.9343540	-2.733077e+00	-2.2888957	15.6053940	5.1618907
MENOAGE	0.5785717	-9.798184e-05	0.4398336	-1.6403031	-0.6281555
REGLAGE	0.3683804	-2.919594e-01	0.1214206	0.7997876	-0.1387678
fruitslegumes	4.7051716	3.375790e+00	-3.6513906	2.0976262	6.3663674
aphyq	2.5208719	-1.533750e+00	-0.6521197	3.9192654	1.6905482
niv_etude	-0.6288470	-7.059056e-02	1.8637694	-4.7071493	-3.7593858
tabac	6.9374871	7.254578e+00	-0.7942813	0.8336979	-1.2341048
menopause	28.3923409	2.013827e+00	-11.2825477	49.1934572	20.2979257

On peut donc faire un tableau récapitulatif de ce qui ressort dans chaque cluster. On ne va pas mettre les maladies courantes qui se retrouvent dans chaque cluster pour faire ressortir les maladies plus rares et pouvoir dégager des facteurs qui les engendrent. On ne met donc pas l'hypertension, l'ostéoporose, les rhumato et la migraine. On met les maladies seulement dans les clusters où elles ont leur plus fort taux de présence. De même pour les facteurs, on met dans le tableau le facteur lorsqu'il a son augmentation ou sa diminution la plus forte pour dégager les facteurs qui ont un rôle important dans le cluster. On fait bien attention que s'exposer au soleil signifie avoir une valeur proche de 1, alors que ne pas s'exposer une valeur proche de 3. Donc une augmentation de la variable soleil correspond à une diminution de l'exposition au soleil. Le nombre de + ou de - dans le tableau représente si le facteur est très supérieur/inférieur à la moyenne de toutes les femmes de l'étude.

/	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
nbre personne	1084	9913	59235	1213	27550
maladies plus présentes	++ kcr diab / / /	++ cas_pso + polyar ksein kpeau /	kseins kpeau / / /	+ valid_AVC + MI ksein kpeau diab	ksein kpeau asthme diab /
maladies moins présentes	toutes sauf kcr /	MI/AVC diabète/kcr	toutes sauf les cancers /	/ /	AVC/MI pso/poly
facteurs +	+ age age1ergross + fruitslegumes aphyq + tabac ++ menopause + soleil	+ tabac age1ergross fruitslegumes / / / /	age1ergross niv_etude / / / / /	imc + age + nbenfviv ++ allaitement + aphyq +++ menopause soleil	imc soleil + allaitement + fruitslegumes ++ menopause / /
facteurs -	/ / / / /	soleil allaitement aphy / /	imcq age allaitement - fruitslegumes – menopause	- encouple menoage – niv_etude / /	soleil niv_etude / / /

A partir de ce tableau on peut déduire des tendances:
- Un IMC trop élevé par rapport à la normale semble entrainer du diabète. Si de plus il est accompagné d'un âge élevé, de la menopause, d'une durée d'allaitement élevée et d'un nombre d'enfant plus élevé que la moyenne, il peut entrainer des AVC et des infarctus (MI). Tout cela semble plutôt cohérent avec les connaissances actuelles de ces maladies.
- Une augmentation de l'IMC est liée à une diminution du niveau d'étude.
- Une première grossesse tardive combinée à la consommation de tabac semble entrainer plusieurs maladies (notamment lorsqu'elle est accompagnée d'autres facteurs). Ces maladies sont le cancer colorectal, le cancer du sein, le cancer de la peau, le psoriasis et la polyarthrite.
- La plupart des maladies ont l'air d'avoir plus de chance d'apparaître lorsque les femmes sont ménopausées. Seulement, on ne peut pas dire si cela est lié à la ménopause en elle-même, ou à l'âge des femmes qui sont souvent plus âgées si elles sont ménopausées.
- Le fait de ne pas être en couple semble favoriser les infarctus et les AVC.
- L'IMC, le tabac, la ménopause et l'allaitement semblent être les facteurs les plus discriminants pour avoir ou non une maladie.

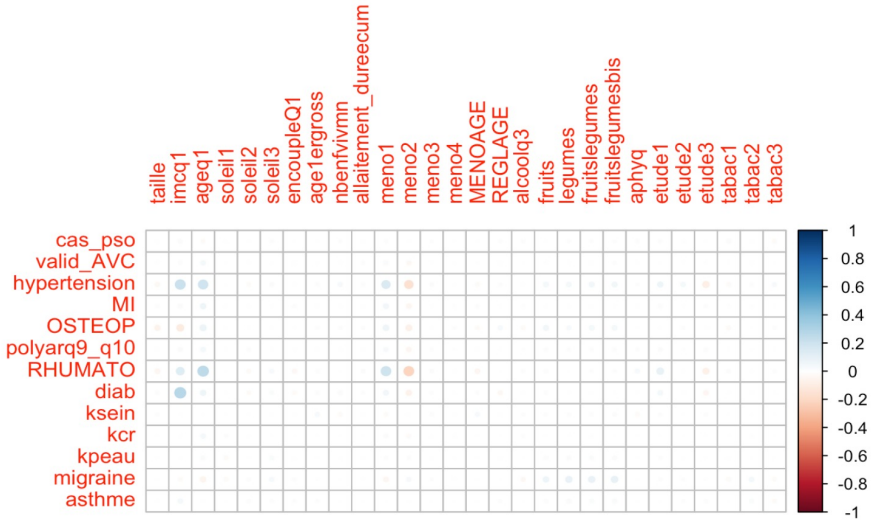
De plus, on peut comparer les maladies dans chaque cluster avec les clusters de maladies réalisés dans la partie précédente.
- On retrouve le cancer colorectal assez isolé des autres maladies, comme trouvé précédemment.
- Les deux autres cancers (de la peau et du sein) semblent aussi faire partie d'un même groupe de maladies. On avait de même trouvé ce résultat précédemment.
- Le psoriasis et la polyarthrite semblent faire partie d'un même cluster. De même, ce résultat est cohérent avec ce qui a été trouvé avant.
- L'AVC, et l'infractus semblent aussi faire partie du même cluster. De même cela est cohérent avec la partie précédente.
- Avec ce tableau, on ne peut rien dire du diabète, de l'hypertension, l'ostéoporose, les rhumatismes, la migraine et l'asthme.

Cependant, ce tableau ne montre pas des liens que l'on pensait évident comme par exemple

qu'une forte exposition au soleil entraine un cancer de la peau. On peut penser que cela est du au faible nombre de femme ayant un cancer de la peau dans cette étude. De même, la consommation de fruis et légumes ne semble pas tellement influencer les maladies, alors que l'on pourrait penser qu'une bonne alimentation empêche l'apparition de maladie.

4.3 Lien entre les facteurs et les maladies

Figure 11: Matrice des corrélations entre maladies et facteurs d'exposition

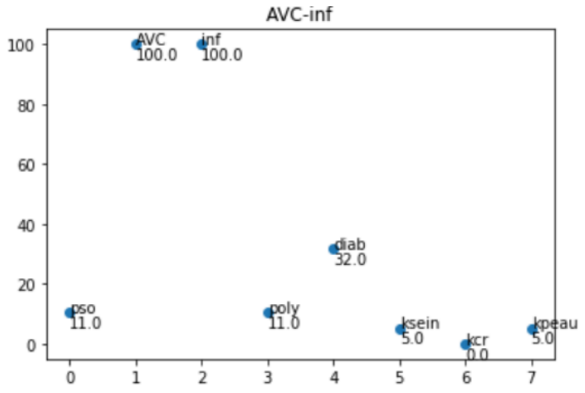


La figure 11 montre la corrélation entre maladies et facteurs. On remarque que l'IMC, l'âge et le fait d'être ménopausée ou non (qui est un facteur lié à l'âge) sont très corrélés (négativement ou positivement) avec certains événements de santé.

5 Cohérence des résultats trouvés

5.1 Cohérence du clustering

Figure 12: Probabilité du trio AVC/infarctus/diabète

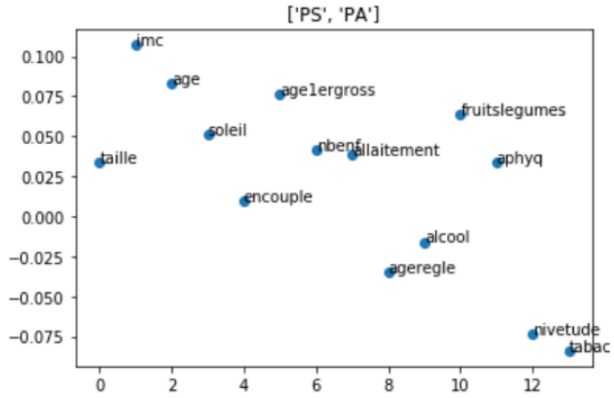


On voit dans ce schéma que la probabilité d'avoir du diabète en ayant déjà eu un AVC

et un infarctus est de 32%, ce qui est haut. Cela corrobore le fait que l'AVC, l'infarctus et le diabète soient dans le même cluster, comme trouvé avec le dendrogramme. On retrouve de même que le psoriasis et la polyarthrite, ainsi que les cancers du sein et de la peau sont souvent associés.

5.2 Cohérence du tableau récapitulatif

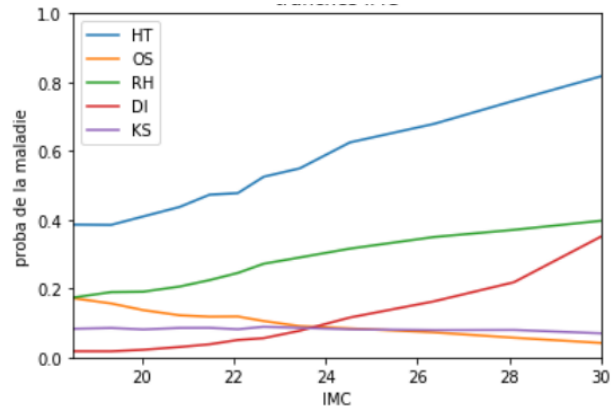
Figure 13: Facteurs influents du cluster PA/PS



On étudie l'influence des variables d'exposition (ici centrées réduites) pour les différents clusters trouvés grâce au dendrogramme. Cette figure nous permet non seulement de voir l'écart à la moyenne des différents paramètres mais aussi les écarts entre les différents facteurs d'exposition. Par exemple, pour le duo polyarthrite et psoriasis on retrouve certains facteurs influents mentionnés dans le tableau récapitulatif : le tabac (dans ce schéma avoir une moyenne négative veut dire fumer plus que la moyenne car non-fumeur = 3) ; l'âge de la première grossesse et les fruits et légumes qui sont ici plus élevés que la moyenne.

5.3 Zoom sur l'IMC

Figure 14: Variation de l'IMC par maladie



Dans ce schéma nous n'avons gardé que les maladies dépendantes de l'IMC. On peut ainsi voir qu'un fort IMC augmente les chances d'avoir de l'hypertension et du diabète,

les deux maladies pour lesquelles on voit une forte variation. A l'inverse, l'ostéoporose atteint davantage les gens en sous-poids. On peut ainsi voir l'influence de ce paramètre sur diverses maladies.

5.4 Comparaison des différentes méthodes de corrélation en Python

Afin d'étudier l'influence des variables d'exposition sur une seule maladie, on a utilisé deux manières différentes :

Figure 15: Variables d'exposition pour l'AVC - Matplotlib

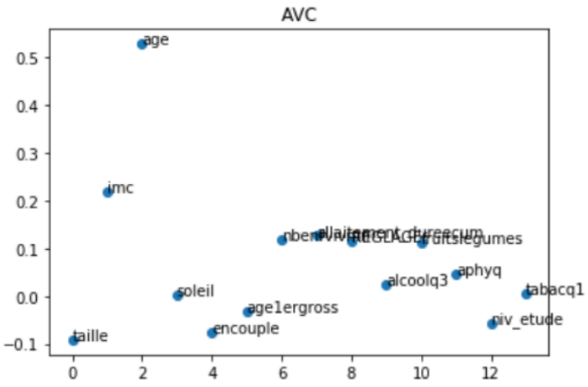
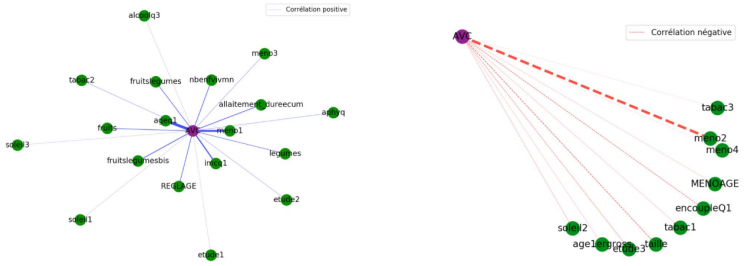


Figure 16: Variables d'exposition pour l'AVC - NetworkX



Avec matplotlib on a mesuré la moyenne de chaque variable centrée réduite parmi les malades d'AVC. Sur NetworkX on trace la corrélation entre les variables. Pour plus de lisibilité on a séparé les corrélations négatives et positives.
Avec ces deux méthodes on trouve des résultats identiques : l'âge et l'IMC très hauts, avec certains paramètres hormonaux également.

6 Conclusion

Durant cet enseignement d'intégration nous avons essayé de nous approprier au mieux la base de données proposée. Nous avons essayé de dégager des patterns de maladies. Grâce aux multiples méthodes de visualisation que nous avons utilisées (dendrogramme, ACP, matplotlib, NetworkX...), nous avons pu en dégager plusieurs. Parmi les clusters que nous avons trouvés, celui qui nous semble le plus pertinent est celui qui rassemble l'AVC, l'infarctus et le diabète. On a pu analysé les facteurs qui influencent fortement ce cluster, tels que l'IMC et l'âge. La richesse de la base de données a pu nous permettre d'étudier

ces facteurs en détail en essayant d'analyser du mieux possible les résultats visuels ou numériques.

References

[1] Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991.

[2] Kassambara. Articles - méthodes des composantes principales dans r: Guide pratique, october 2017. http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/73-acp-analyse-en-composantes-principales-avec-r-l-essentiel/.

A Annexes

A.1 Gestion des données

Google Colab utilisé pour nettoyer la base de données : `https://colab.research.google.com/drive/1FvFODLFHhOMHdgW1v9YfALyk5ZtY_biS?usp=sharing`

A.2 Code R pour le CAH des maladies

```
library(ClustOfVar)
# importation des donnes
X <- read.csv("base_E3N_clean.csv", header = TRUE, sep =",",
stringsAsFactors = TRUE)

# On cree une seule colonne pour la migraine : 1 si cas de migraine
(ancien ou actuel), 0 si non cas ; idem pour l'asthme

X$migraine <- X$migraine1 + X$migraine2
X$asthme <- X$asthme2 + X$asthme3

# On realise le clustering avec la biblioth que ClustOfVar

arbre <- hclustvar(X.quanti = X[c(39, 44:55)])
plot(arbre)
```

A.3 Code R pour l'ACP et le k-means de la partie 4.2

```
library(FactoMineR)
library(factoextra)
library(ggplot2)
library(corrplot)

#extraction des donnes
base_E3N <- read.csv("E:/Documents/ETUDES/CentraleSupelec/1A/
base_E3N_clean.csv", header = TRUE, sep = ",", stringsAsFactors = TRUE)

X = base_E3N[, 3:NCOL(base_E3N)]
X$migraine <- X$migraine1 + X$migraine2
X$asthme <- X$asthme2 + X$asthme3
X$tabac <- X$tabac1 + X$tabac2
X$menopause <- X$meno1 + X$meno3
Y = X[,c(37, 42:53)]

#ACP
res.pca <- PCA(Y, scale.unit = TRUE, ncp = 6, graph = FALSE) #realise l ACP
print(res.pca)
get_eigenvalue(res.pca) #donne les valeurs propres
fviz_pca_biplot(res.pca, col.var = "#2E9FDF",col.ind = "#696969",
label = "var") #trace le biplot
```

```
fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 50)) #trace le
pourentage expliqu  de variance par chaque valeur propre

# Qualite de representation cos2
corrplot(var$cos2, is.corr=FALSE)
head(var$cos2, 13)
fviz_pca_var(res.pca, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE #   vite   le chevauchement de texte
)

#contribution des variables sur chaque dimension
corrplot(var$contrib, is.corr=FALSE)
fviz_pca_var(res.pca, col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07")
)
head(var$contrib, 13)

## Coordonnees des variables
corrplot(res.pca$var$coord, is.corr=FALSE,method="number")

ind <- get_pca_ind(res.pca)

# Colorer les indiv d'un cluster
#(cluster fait sur les coordonn es des points)

groupes.kmeans <- kmeans(ind$coord,centers=5,nstart=100)
print(groupes.kmeans)
grp <- as.factor(groupes.kmeans$cluster)
fviz_pca_biplot (res.pca,
                 col.ind = grp, palette =
                 c("#00AFBB", "#E7B800","#EA0034","#B610BF","#9AF000"),
                 addEllipses = TRUE, label = "var",
                 col.var = "black", repel = TRUE,
                 legend.title = "Cluster")

inertie.expl <- rep(0,times=10)
for (k in 2:10){
  clus <- kmeans(ind$coord,centers=k,nstart=10)
  inertie.expl[k] <- clus$betweenss/clus$totss
}
#graphique
plot(1:10,inertie.expl,type="b",xlab="Nb. de groupes",
ylab="% inertie expliqu e")

### Recuperer les maladies de chaque cluester
str(Y)
taille_cluster <- groupes.kmeans$size
```

```
#pour le groupe 1
cas_pso1 <- 0
for (k in 1:NROW(Y)){
  if (grp[k] == 1){
    cas_pso1 <- cas_pso1+Y$cas_pso[k]
  }
}
cas_pso1 <- cas_pso1/taille_cluster[1] *100
#ca sera le m me code pour chaque maladie et chaque cluster


#### FACTEURS PAR CLUSTER : ecart relatif a la moyenne

# Pour cluster 1

imcq11 <- 0
for (k in 1:NROW(Y)){
  if (grp[k] == 1){
    imcq11 <- imcq11+X$imcq1[k]
  }
}
imcq11 <- (imcq11/taille_cluster[1] - mean(X$imcq1))/mean(X$imcq1) *100
#ca sera la m me chose pour les autres facteurs de chaque cluster,
juste il fallait faire attention quand il y avait encore des valeurs
N.A comme ici :
age1ergross1 <- 0
nb_age1 <- 0
for (k in 1:NROW(Y)){
  if (grp[k] == 1 & !is.na(X$age1ergross[k])){
    age1ergross1 <- age1ergross1+X$age1ergross[k]
    nb_age1 <- nb_age1 +1
  }
}
age1ergross1 <- (age1ergross1/nb_age1 - 24.0)/24.0 *100
```

A.4 Figures

A.4.1 Visualisation par des graphes

A.4.2 Figures supplémentaires pour l'ACP

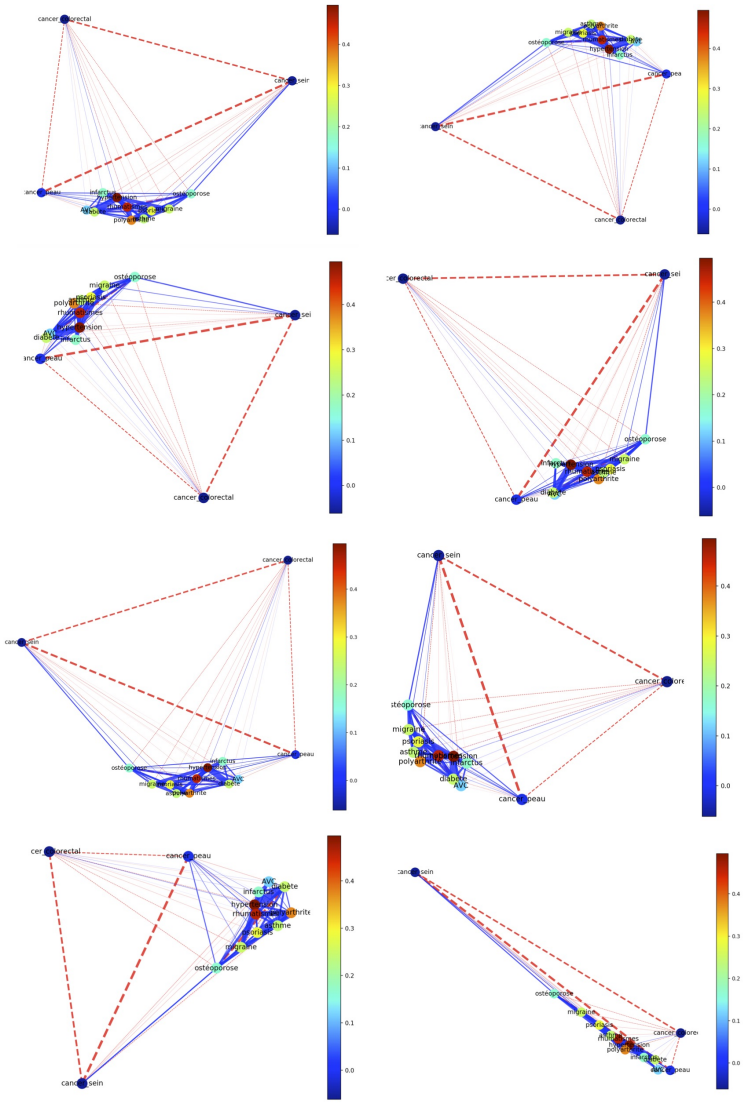


Figure 17: 8 graphes différents représentés
L'algorithme de Fruchterman-Reingold minimise l'énergie du système, mais il peut arriver de se retrouver dans un minimum local, c'est pourquoi nous avons exécuté le code plusieurs fois pour s'assurer de la validité du résultat. Les 7 premiers graphiques sont pratiquement identiques, seul le dernier semble un peu particulier : il s'agit probablement d'un minimum local.

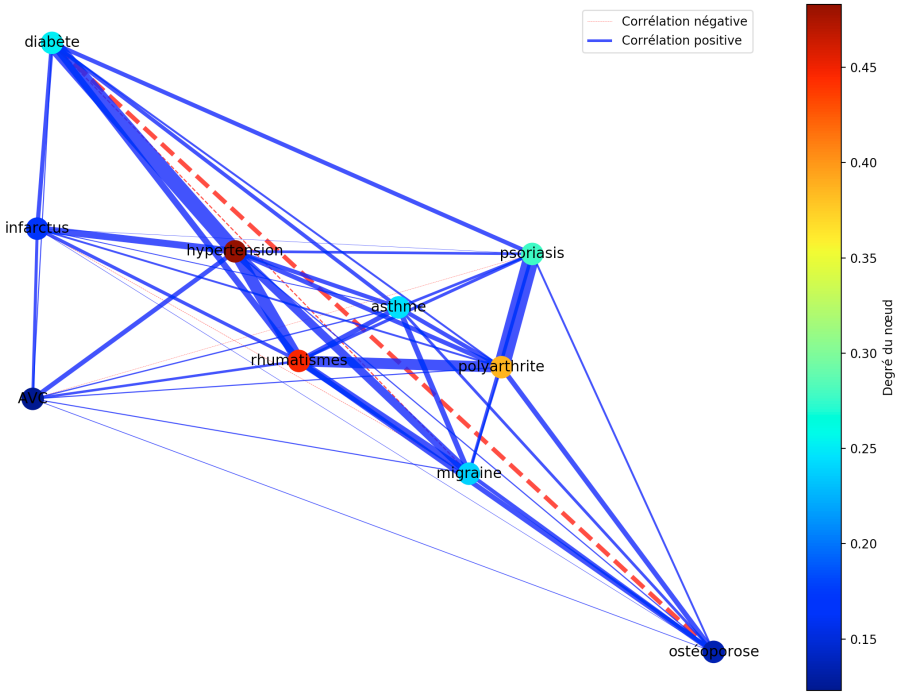


Figure 18: Pour plus de visibilité, nous avons également tracé un graphe de corrélation des maladies excluant les cancers.
On remarque une corrélation importante entre diabète et hypertension ou psoriasis et polyarthrite par exemple.

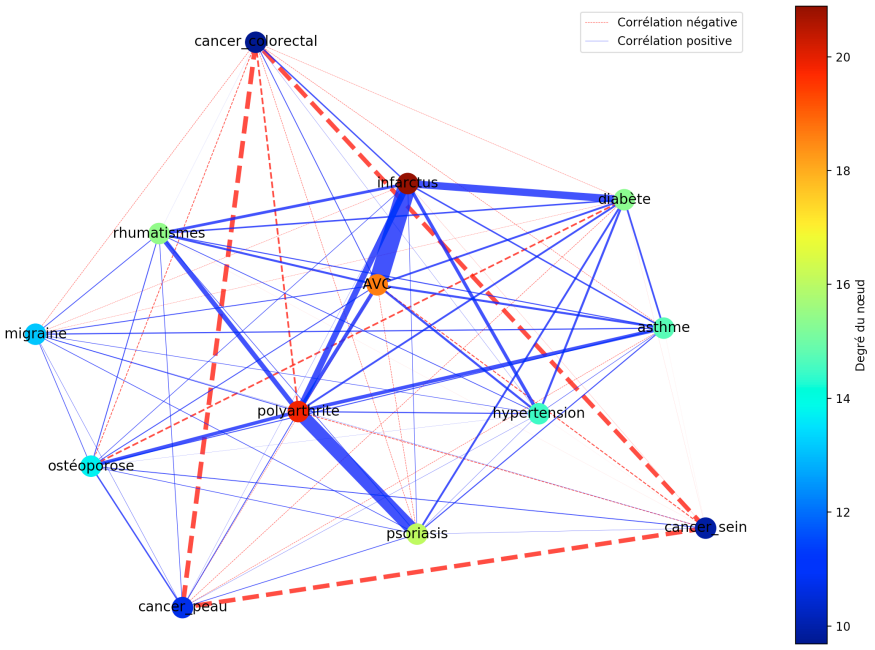


Figure 19: Les graphiques précédents utilisaient le coefficient de corrélation comme définition du poids de l'arc. Nous avons également expérimenté avec d'autres façons de définir le poids. Ici, $w_{AB} = \frac{P(A \cap B)}{P(A)P(B)}$.
Nous obtenons des résultats cohérents, par exemple une corrélation importante entre AVC et infarctus, entre polyarthrite et psoriasis, ainsi qu'une décorrélation des cancers avec la plupart des maladies.

Figure 20: Contribution de chaque variable par dimensions

```
> head(var$contrib, 13)
                    Dim.1       Dim.2       Dim.3       Dim.4       Dim.5                    Dim.6
cas_pso      9.731183e+00 14.82327346 19.506659675  4.84074872  0.007261827 cas_pso      3.996944e-04
valid_AVC    1.794479e+00  2.95204440  2.348015137  1.91882112  1.414121115 valid_AVC    3.320093e+01
hypertension 2.176552e+01 13.89340560  1.138381625  0.03449406  0.290597237 hypertension 7.198456e-01
MI           3.739622e+00  8.54168972  0.009390916  0.01693223  0.452069045 MI           2.891095e+01
OSTEOP       2.331787e+00 16.25730773 12.932653657  3.53102850  0.122583385 OSTEOP       7.792735e+00
polyarq9_q10 1.558220e+01 16.78125062 12.296947395  0.91721437  0.211341522 polyarq9_q10 2.457235e+00
RHUMATO      2.041495e+01  0.02137687  1.532017077  1.39802293  0.138209725 RHUMATO      1.277359e-02
diab         1.068624e+01 21.42931253  4.603737002  5.34196162  2.478538026 diab         4.564469e+00
ksein        1.782183e-04  1.26718753 15.889108828 38.03475701  1.154900843 ksein        1.103856e+00
kcr          1.987248e-02  1.58104182  3.918363222  4.60949035 62.185092327 kcr          1.905199e-02
kpeau        2.182465e-01  0.44476824  1.636737169 36.43151361 27.535223669 kpeau        1.927724e-01
migraine     7.177027e+00  1.61792545 20.394015116  2.90620590  1.067120670 migraine     1.149411e+01
asthme       6.538702e+00  0.38941602  3.793973180  0.01880958  2.942940610 asthme       9.530871e+00
```

Figure 21: Pourcentage d'inertie expliquée (k-means sur les coordonnées des données de l'ACP

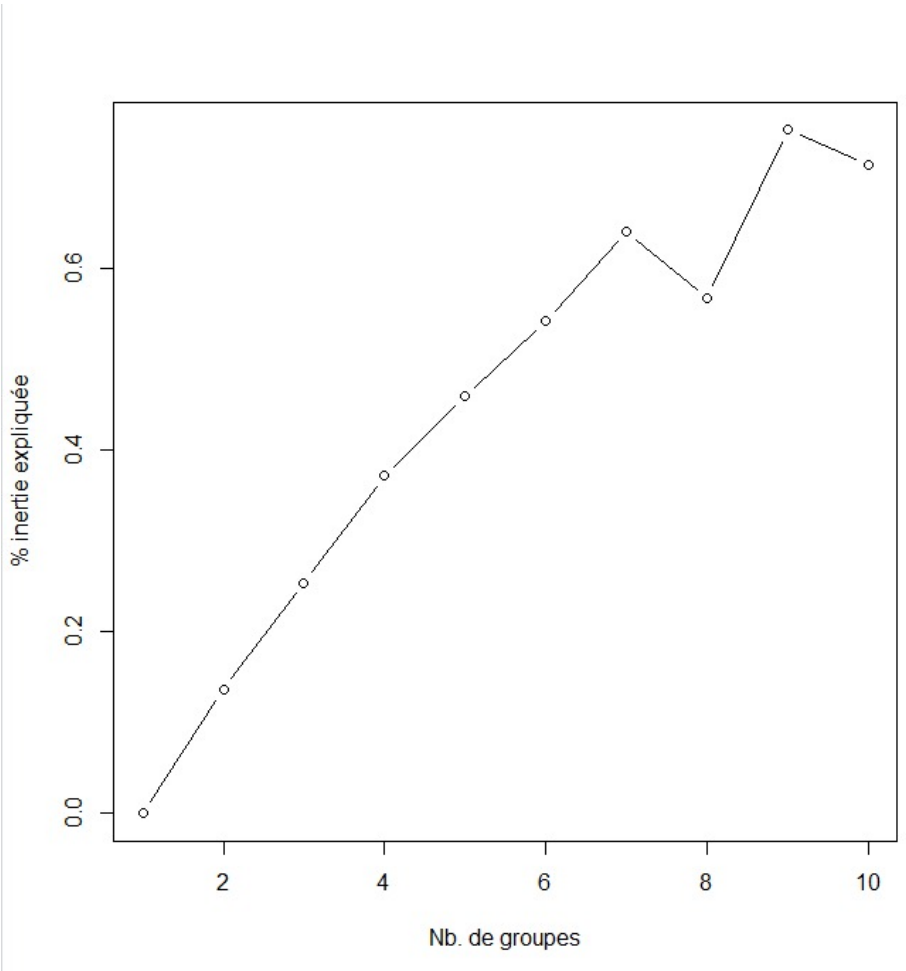


Figure 22: Biplot de l'ACP avec 9 clusters fait par k-means