L'analyse en composantes principales comme outil biostatistique : une routine pour étudier une structure de biomarqueurs

par

Francis DUSSEAULT-BÉLANGER

mémoire présenté au Département de mathématiques
en vue de l'obtention du grade de
maître ès sciences (M.Sc.)

FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, mars 2013

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-95165-1

Our file Notre référence
ISBN: 978-0-494-95165-1

Le 27 mars 2013

le jury a accepté le mémoire de Monsieur Francis Dusseault-Bélanger dans sa version finale.

Membres du jury

Professeur Ernest Monga
Directeur de recherche
Département de mathématiques

Professeur Alan Cohen
Codirecteur de recherche
Faculté de médecine et des sciences de la santé

Professeure Marie-France Hivert
Évaluateur interne au programme
Faculté de médecine et des sciences de la santé

Professeur Taoufik Bouezmarni
Président rapporteur
Département de mathématiques

SOMMAIRE

En statistiques, l'analyse en composantes principales (ACP) est une technique couramment utilisée afin de détecter la présence de processus gérant une base de données, c'est-à-dire des principaux axes de variation des données. Toutefois, on se contente trop souvent d'exécuter l'algorithme et d'interpréter directement les résultats sans analyses postérieures concernant la stabilité ou la généralisation des résultats.

Pourtant, comme toute mesure statistique prélevée sur un échantillon, l'ACP peut présenter des résultats spécifiques à l'échantillon et difficilement rendre compte de la population étudiée. Il est donc important de développer des analyses complémentaires nous permettant d'évaluer la stabilité des résultats de l'ACP et de déterminer si, oui ou non, les résultats tirés de l'échantillon sont dignes de confiance.

On construira et mettra cette nouvelle routine à l'épreuve en étudiant deux sujets distincts tant par leur nature que par leur complexité. On étudie en un premier temps le syndrome métabolique, syndrome considéré comme un facteur important dans le développement du diabète et des maladies cardiovasculaires. Ensuite, on étudie un phénomène beaucoup plus complexe et surtout moins bien défini que le syndrome métabolique, le processus physiologique du vieillissement.

Mots-clés : Analyse en composantes principales, biomarqueurs, syndrome métabolique, statistique, vieillissement.

À toutes celles et ceux
qui ont cru en moi.

REMERCIEMENTS

Je tiens d'abord à remercier Pr. Alan Cohen pour sa patience et son dévouement tout au long de nos travaux de recherche, plus particulièrement lors des derniers milles lorsque je partageais mon temps entre l'Université de Sherbrooke et le cégep de Drummondville ainsi que Pr. Ernest Monga pour son appui lors de mon parcours académique au début comme à la fin de mes études.

Je veux aussi remercier tous les membres de l'équipe PRIMUS au Centre de Recherche Clinique Étienne-Le-Bel qui m'ont assisté de près ou de loin pendant mes travaux de recherche. Je ne peux non plus passer sous le silence la contribution de la Pre Marie-France Hivert en tant que mentor lors de mon dernier stage coopératif et comme collègue pendant la maîtrise.

Enfin, je tiens à souligner l'appui de mes proches qui m'ont appuyé durant ce beau périple qu'est l'obtention d'un diplôme de 2e cycle en sciences. Plus particulièrement, ma copine Catherine pour m'avoir appuyé dans les hauts comme dans les bas.

Avant de terminer, soulignons aussi l'apport financier considérable de l'axe ARIES du Centre Hospitalier Universitaire de Sherbrooke (CHUS) ainsi que de l'Institut des Sciences Mathématiques (ISM).

Mes plus sincères remerciements à toutes et à tous,

Francis Dusseault-Bélanger

Sherbrooke, décembre 2012

Table des matières

SOMMAIRE iii

REMERCIEMENTS iv

TABLE DES MATIÈRES v

LISTE DES FIGURES vii

Chapitre 1 — Mise en contexte 1

1.1 L'analyse en composantes principales 1

1.2 Le syndrome métabolique 3

1.3 Le vieillissement 3

1.4 Objectifs 4

Chapitre 2 — Méthodologie 5

2.1 L'analyse en composantes principales 6

2.1.1 En théorie 6

2.1.2 En pratique 9

2.1.3 En images 10

2.2 L'analyse parallèle 11

2.3 L'algorithme de Daudin, Duby et Trecourt 13

2.4 Les scores alternatifs 15

Chapitre 3 — Le syndrome métabolique 16

Chapitre 4 — Le processus du vieillissement 36

Chapitre 5 — CONCLUSION 58

5.1 Atteinte des objectifs 58

5.1.1 Le premier objectif 58

5.1.2 Le second objectif 59

5.1.3 Le troisième objectif 59

5.2 Avenues de recherche futures 60

5.2.1 Concernant le syndrome métabolique 60

5.2.2 Concernant le vieillissement 60

BIBLIOGRAPHIE 69

Liste des figures

2.1 Illustration de l'ACP 10

2.2 Illustration de l'anàlyse parallèle de Horn 12

2.3 Illustration des résultats de l'algorithme de Daudin, Duby et Trecourt 14

Chapitre 1

Mise en contexte

1.1 L'analyse en composantes principales

Il est difficile de retracer précisément les origines de l'ACP. Par exemple, Jolliffe [11] note que Beltrami [2] et Jordan [12] ont développé indépendamment la décomposition en valeurs singulières dans une forme qui ressemble à celle de l'ACP entre 1873 et 1874.

Un peu plus tard au 20^e siècle, Fisher et Mackenzie [7] utilisent cette même décomposition dans un contexte d'analyse bidimensionnelle durant une expérience agricole en 1923 qui ressemble à celle d'une ACP.

Toutefois, la communauté scientifique accorde plutôt les fondements de l'ACP à Pearson (1901) [14] et Hotelling (1933) [10] même si les deux mathématiciens adoptent des approches différentes. Hotelling préfère une approche plus algébrique alors que Pearson cherche à trouver la droite ou le plan s'ajustant le mieux à un ensemble de points de grande dimension se rapprochant ainsi un plus plus d'une méthode statistique.

Il est intéressant de noter jusqu'à quel point Pearson a réussi à développer la théorie de l'ACP malgré la quasi-inexistence de méthodes numériques. Il mentionne même dans son article que l'ACP peut facilement être appliquée à des problèmes numériques, mais que les calculs deviennent lourds après la quatrième ou cinquième dimension.

Quant à lui, Hotelling établit les bases de l'ACP en mentionnant pour la première fois le terme *composante principale* et en les classant de telle sorte qu'elles maximisent successivement la variance expliquée par les variables originales. De plus, Hotelling met en garde ses lecteurs de bien distinguer l'analyse en composantes principales et l'analyse factorielle développée un peu plus tôt dans la même décennie puisque les deux méthodes sont fondamentalement différentes même si elles peuvent sembler similaires en surface puisque l'analyse factorielle utilise des méthodes se rapprochant des tests d'hypothèse alors que l'ACP est une méthode plus proche des statistiques descriptives.

En résumé, l'idée centrale de l'analyse en composantes principales (ACP) est de réduire la dimension d'un ensemble de données illustrant plusieurs variables plus ou moins reliées entre elles en retenant le plus de variance expliquée possible. Ceci nous permet de découvrir et interpréter les liens unissant les variables étudiées.

Ce processus est rendu possible en transformant les données originales en un tout nouvel ensemble de variables ordonnées de telle sorte qu'elles expliquent successivement un maximum de variance présente dans les variables originales.

Ainsi, l'ACP telle que définie par Pearson et formalisée par Hotelling nous permet de répondre à deux questions : (1) Peut-on réduire la dimension de la structure de données ? (2) Si on le peut, comment peut-on résumer la complexité de la structure de données avec une perte d'information (ou de variance expliquée) minimale ?

Autrement dit, l'ACP nous permet de détecter la présence d'une structure de corrélation, c'est-à-dire des principaux axes de variation, émergeant d'une structure de données. C'est pourquoi l'ACP est l'outil privilégié pour valider un sujet tel que le syndrome métabolique ou bien étudier un concept mal défini tel que le processus physiologique du vieillissement puisqu'elle n'exige aucune hypothèse de base et néglige ainsi les risques de biais.

1.2 Le syndrome métabolique

Depuis plusieurs années déjà, plusieurs experts, dont Padwal et Sharma [4], ont relevé l'importance de mesures préventives pour réduire les risques de diabète et de maladies cardiovasculaires. On s'intéresse à une de ces mesures en particulier, le syndrome métabolique.

Le syndrome métabolique est une mesure de paramètres cliniques facilement et couramment mesurés. En fait, un patient est atteint du syndrome s'il remplit au moins trois des cinq critères suivants :

1. Un surplus de poids
2. Une tension artérielle élevée
3. Un taux de sucre sanguin élevé
4. Un taux de cholestérol HDL bas
5. Un taux de triglycérides élevé

Plusieurs médecins chercheurs étudient encore en détail le syndrome métabolique pour évaluer son potentiel de prévention des maladies cardiovasculaires et du diabète. Toutefois, même s'il est bien défini théoriquement, le syndrome demeure à ce jour un concept qui reste à confirmer en pratique : est-ce un réel syndrome ou une réunion de facteurs indépendants ? C'est ce que nous tenterons de démystifier un peu plus tard.

1.3 Le vieillissement

En jetant un coup d'oeil à l'histoire récente, on remarque aussi que l'étude du vieillissement et le développement du domaine de la gérontologie se sont développés à une vitesse impressionnante. Plusieurs domaines des sciences ont même contribué à son développement, de l'épidémiologie à la démographie en passant évidemment par les statistiques.

Malgré tout, la recherche des causes du vieillissement est encore d'actualité : on ne connaît toujours pas ce qui engendre le processus physiologique du vieillissement. En d'autres mots, on se demande encore pourquoi l'être humain vieillit.

Notre propre théorie consiste à considérer le processus physiologique du vieillissement comme une perturbation de la régulation des biomarqueurs humains (homéostasie).

1.4 Objectifs

Dans les faits, on désire atteindre spécifiquement trois objectifs en effectuant cette recherche :

1. Développer une routine d'analyses nous permettant d'analyser en profondeur une base de données à l'aide de l'ACP. Cette routine devra nous permettre de détecter une structure de corrélation et d'identifier si cette structure est représentative de la population dans son intégrité.
2. Utiliser la routine élaborée pour confirmer le bien-fondé de la définition du syndrome métabolique à travers des données réelles recueillies via *Ariane*, la base de données cliniques du *Centre Hospitalier Universitaire de Sherbrooke* (CHUS).
3. Utiliser la routine élaborée afin d'identifier si la structure de corrélation sous-jacente aux données du *Women's Health and Aging Study* (WHAS) s'apparente au processus physiologique du vieillissement.

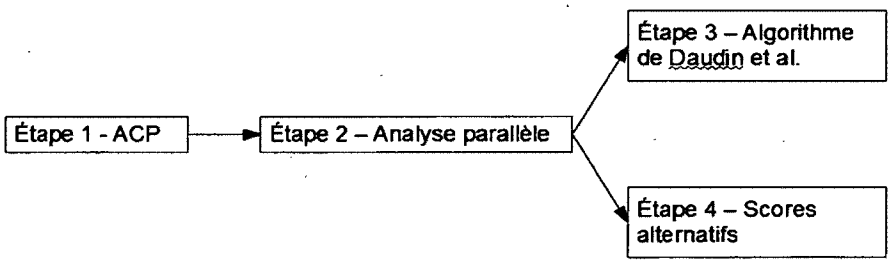
Chapitre 2

Méthodologie

La routine développée se résume en quatre points :

1. Effectuer l'ACP sur les données.
2. Déterminer la dimension de la structure de données, c'est-à-dire choisir le nombre de nouvelles variables à retenir (Analyse parallèle de Horn).
3. Analyser la stabilité de la structure de données soumise à un rééchantillonnage *bootstrap* (Algorithme de Daudin, Duby et Trecourt).
4. Analyser la stabilité des résultats à travers diverses sous-populations (Scores alternatifs).

On fait un bref survol des détails techniques de chacune des étapes dans les sections qui suivent.



2.1 L'analyse en composantes principales

On fait un bref survol des notations utilisées dans les sections suivantes :

1. Lettre majuscule en gras ($\mathbf{A}$) $\rightarrow$ vecteur aléatoire
2. Lettre majuscule (A) $\rightarrow$ variable aléatoire
3. Lettre minuscule en gras ($\mathbf{a}$) $\rightarrow$ vecteur d'observations d'une variable aléatoire A
4. Lettre minuscule (a) $\rightarrow$ observation d'une variable aléatoire A
5. Lettre majuscule avec indice $m \times n$ ($A_{m \times n}$) $\rightarrow$ matrice de dimension $m \times n$

2.1.1 En théorie

Soit $\mathbf{X} = (X_1, X_2, ..., X_p)'$ un vecteur aléatoire de dimension p et Σ la matrice de covariance associée, c'est-à-dire

$$\Sigma = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & ... & ... & ... & Cov(X_1, X_p) \\ Cov(X_2, X_1) & Var(X_2) & ... & ... & ... & Cov(X_2, X_p) \\ . & & & & & . \\ . & & & & & . \\ . & & & & & . \\ Cov(X_p, X_1) & Cov(X_p, X_2) & ... & ... & ... & Var(X_p) \end{pmatrix}$$

À moins que p ne soit petit, ou que la structure soit relativement simple, il est souvent difficile d'étudier les $\frac{1}{2}p(p+1)$ variances et covariances unissant les p variables originales.

Une approche alternative est d'étudier un nombre réduit de variables U_i $(i = 1, 2, ..., k)$ $(k < p)$ déduites des variables originales X_i expliquant successivement un maximum de variance expliquée des données originales.

La première étape de l'ACP est donc de trouver une première fonction linéaire $\alpha_1'\mathbf{X}$ expliquant un maximum de variance.

Ensuite, on cherche une fonction linéaire $\alpha_2'\mathbf{X}$ non corrélée avec la précédente qui maximise la variance expliquée restante et ainsi de suite jusqu'à ce qu'on trouve une fonction linéaire $\alpha_p'\mathbf{X}$

non corrélée aux $p-1$ fonctions précédentes. Ainsi, on trouve p vecteurs $\alpha_i' = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{ip})$ tous indépendants entre eux formant une base d'un nouvel espace.

Le processus développé nous permet donc de transformer le vecteur aléatoire $\mathbf{X}$ vers un nouveau vecteur aléatoire

$$\mathbf{U} = (U_1, U_2, ..., U_p)' = (\alpha_1'\mathbf{X}, \alpha_2'\mathbf{X}, ..., \alpha_p'\mathbf{X})'.$$

où $Cov(U_i, U_j) = 0$ $(\forall i = 1, 2, ..., p)(\forall j = 1, 2, ...p)(i \neq j)$
et $Var(U_1) \geq Var(U_2) \geq ... \geq Var(U_p)$.

Toutefois, l'ACP telle que présentée comporte une lacune majeure. En effet, puisqu'on utilise la matrice de covariance Σ, les résultats seront sensibles à l'échelle de mesure utilisée pour chacune des variables.

Par exemple, si on étudie une variable mesurée en centimètres (10^{-2}) et une variable mesurée en kilomètres (10^3), l'ACP accordera beaucoup plus d'importance à la variable mesurée en centimètres puisque sa variance sera numériquement plus élevée que celle de la variable mesurée en kilomètres. On conseille donc de centrer et réduire chacune des variables avant de procéder aux calculs afin de s'assurer que la même échelle de mesure soit utilisée pour l'ensemble des variables.

Sinon, il est de commun usage de substituer la matrice de covariance Σ par la matrice de corrélation pour éviter le problème mentionné plus haut. Par matrice de corrélation, on entend la matrice suivante :

$$R_{p\times p} = \begin{pmatrix} 1 & Corr(X_1, X_2) & ... & ... & ... & Corr(X_1, X_p) \\ Corr(X_2, X_1) & 1 & ... & ... & ... & Corr(X_2, X_p) \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ Corr(X_p, X_1) & Corr(X_p, X_2) & ... & ... & ... & 1 \end{pmatrix}$$

Nous allons maintenant bien définir chacun des concepts impliqués dans l'ACP, c'est-à-dire les *composantes principales*, les *charges* et les *scores*.

Définition. ***Composantes principales***

On appelle composantes principales les fonctions linéaires répondant aux deux critères de maximisation de la variance et d'absence de corrélation. Ce sont en réalité les vecteurs α_i *(*$i = 1, 2, ..., p$*) définis plus haut.*

Définition. ***Charges (ou loadings)***

On appelle charges les coefficients formant les composantes principales. Ce sont en réalité les coefficients α_{ij} *(*$i, j = 1, 2, ..., p$*) des vecteurs* α_i *définis plus haut.*

Définition. ***Scores***

On appelle scores les nouvelles variables aléatoires. Ce sont en réalité les coefficients U_i *(*$i = 1, 2, ..., p$*) du vecteur* $\boldsymbol{U}$ *défini plus haut.*

Maintenant que nous avons défini ce qu'est une composante principale, il est important de savoir comment les calculer.

Soit Σ la matrice de covariance associée au vecteur aléatoire $\mathbf{X}$, alors la i^e composante principale ($i = 1, 2, ..., p$) est le vecteur propre de Σ associé à la valeur propre λ_i où $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ puisque Σ est symétrique et semi-définie positive.

De plus, si α_i est choisi de telle sorte qu'il soit unitaire ($\|\alpha_i\|_2 = 1$), alors la variance de α_i est $Var(\alpha_i) = \lambda_i$. C'est notamment le cas lorsqu'on remplace la matrice de covariance Σ par la matrice de corrélation $R_{p \times p}$.

En résumé, l'ACP obéit à deux importantes contraintes : (1) Les composantes principales doivent expliquer successivement un maximum de variance et (2) elles doivent toutes être non corrélées entre elles, ce qui entraîne leur indépendance dans le cas de la multinormalité des vecteurs de départ.

Pour une démonstration complète de ces deux propriétés, on se réfère au besoin au livre de Jolliffe [11].

2.1.2 En pratique

En pratique, nous travaillons rarement sur un vecteur aléatoire. On travaille plutôt sur des matrices d'obsèrvations regroupant n observations sur p variables, c'est-à-dire

$$X_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \dots & \dots & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & \dots & \dots & x_{2p} \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ x_{n1} & x_{n2} & \dots & \dots & \dots & x_{np} \end{pmatrix}$$

Ainsi, on utilise la matrice $S_{p\times p}$ comme matrice de corrélations, c'est-à-dire :

$$S_{p\times p} = \begin{pmatrix} Var(x_{.1}) & Cov(x_{.1}, x_{.2}) & \dots & \dots & \dots & Cov(x_{.1}, x_{.p}) \\ Cov(x_{.2}, x_{.1}) & Var(x_{.2}) & \dots & \dots & \dots & Cov(x_{.2}, x_{.p}) \\ . & & & . & & . \\ . & & & & . & . \\ . & & & & . & . \\ Cov(x_{.p}, x_{.1}) & Cov(x_{.p}, x_{.2}) & \dots & \dots & \dots & Var(x_{.p}) \end{pmatrix}$$

où $x_{.j}$ représente le j^e colonne de $X_{n\times p}$.

On extrait ensuite les vecteurs propres de la matrice $S_{p\times p}$ pour former la matrice $A_{p\times p}$ regroupant chacune des composantes principales comme colonne, c'est-à-dire :

$$A_{p\times p} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \dots & \dots & \alpha_{1p} \\ \alpha_{21} & \alpha_{22} & \dots & \dots & \dots & \alpha_{2p} \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ \alpha_{p1} & \alpha_{p2} & \dots & \dots & \dots & \alpha_{pp} \end{pmatrix}$$

où α_{ij} représente la i^e charge de la j^e composante principale.

On trouve enfin la matrice des scores , c'est-à-dire :

$$U_{n\times p} = X_{n\times p} \times A_{p\times p} = \begin{pmatrix} u_{11} & u_{12} & \dots & \dots & \dots & u_{1p} \\ u_{21} & u_{22} & \dots & \dots & \dots & u_{2p} \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ u_{n1} & u_{n2} & \dots & \dots & \dots & u_{np} \end{pmatrix}$$

2.1.3 En images

On peut imaginer l'ACP comme une illustration toute simple. Supposons qu'on dispose d'un objet tridimensionnel (une statue par exemple) au centre d'une pièce.

Sommes-nous en mesure de reconstituer une image plus ou moins précise de l'objet sans même regarder l'objet ?

Absolument, il suffit d'effectuer une analyse en composantes principales : les variables originales seront les différentes formes de complexité représentées par l'objet et les composantes principales seront les ombres que l'objet projette sur les murs.

Ainsi, en supposant qu'on peut observer les ombres, il sera possible de reconstituer l'objet au centre de la pièce en regardant trois ombres projetées sur les murs de la pièce.

Figure 2.1: Illustration de l'ACP

L'ombre projetée sur le mur nous permet d'avoir une bonne idée de ce que représente la statue. Une seule dimension serait donc suffisante pour reconstruire une réplique plus ou moins fidèle de cette statue.

2.2 L'analyse parallèle

Un problème se pose cependant après avoir effectué l'ACP, combien de composantes principales devons-nous retenir ?

Plusieurs méthodes sont couramment utilisées dans la littérature scientifique, que ce soit le test de Bartlett [1], le test de Kaiser [13], le *scree test* de Cattell [3], l'analyse parallèle de Horn [9] ou toute autre méthode. Toutefois, une d'entre elles a retenu notre attention grâce à sa plus grande fiabilité (selon Zwick et Velicer [8]), soit l'analyse parallèle.

L'approche suggérée par Horn, l'analyse parallèle, s'appuie sur une adaptation du populaire test de Kaiser ($\lambda_i \geq 1$). Horn montre, au niveau de la population, que les valeurs propres de la matrice de corrélation de variables non corrélées sont égales à 1. Toutefois, quand des échantillons sont générés à l'aide d'une telle matrice, les valeurs propres initiales sont plus grandes que 1 alors que les valeurs propres finales sont quant à elles, plus petites que 1.

Horn propose alors de comparer les valeurs propres des données réelles à celles de n ensembles de données aléatoires nòn corrélées. Les composantes principales d'intérêt devraient alors avoir des valeurs propres significativement plus élevées que celles des matrices aléatoires.

Ainsi, il suggère de ne préserver que les k composantes principales correspondant à ce critère.

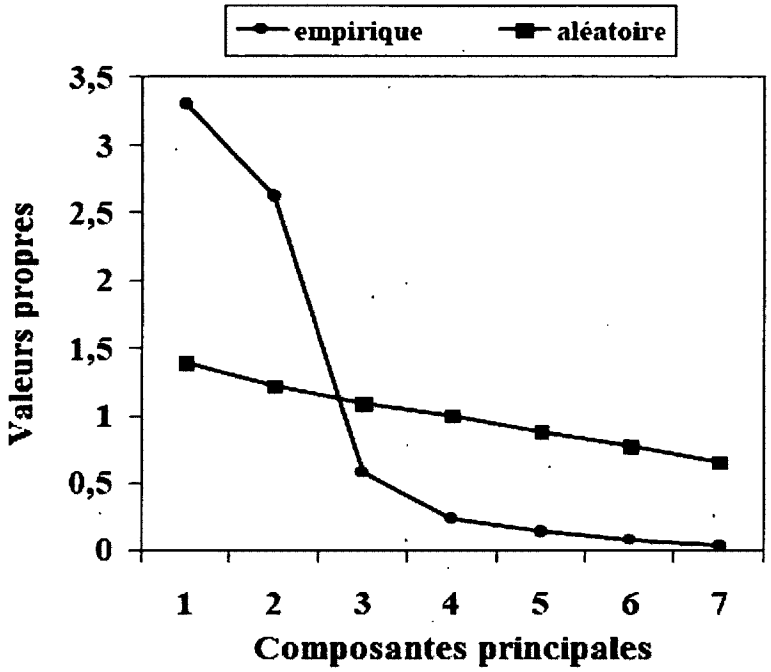
La procédure ressemblerait donc à ceci :

1. Générer les données aléatoires non corrélées n fois
2. Extraire les valeurs propres des n matrices de corrélation aléatoires
3. Calculer la moyenne de chacun des groupes de valeurs propres
4. Comparer les données réelles aux moyennes des données aléatoires

Cette procédure a l'avantage d'intégrer la fiabilité et l'emphase sur la compression des données du test de Kaiser en considérant les effets de la taille de l'échantillon. De plus, un graphique permet facilement de visualiser la procédure (voir figure 1.2).

Pour de plus amples détails, Hayton et al. présente avec brio un résumé détaillé de la procédure dans leur article [6].

Figure 2.2: Illustration de l'analyse parallèle de Horn



Étant donné que seul les deux premières composantes principales présentent des valeurs propres plus grandes que les composantes principales aléatoires, on choisit d'exclure les composantes principales 3 à 7.

2.3 L'algorithme de Daudin, Duby et Trecourt

Le problème étudié par Daudin, Duby et Trecourt [5] consiste à trouver une mesure de variabilité du sous-espace engendré par les k premières composantes principales.

Soit $E_k^{(B)}$ l'espace généré par les k premières composantes principales à la suite d'un *bootstrap* et E_k l'espace généré par les k premières composantes principales de la population. L'équipe de Daudin propose de mesurer la variabilité de $E_k^{(B)}$ autour de E_k à l'aide de la statistique suivante :

$$A_k = \sum_{i,j \leq k} \rho_{ij}$$

où $\rho_{ij} = corr(\mathbf{u}_i, \mathbf{u}_j^{(B)})$ est la corrélation entre le i^e vecteur de scores de la population et le j^e vecteur de scores du rééchantillonnage *bootstrap*.

Ainsi, une stabilité parfaite entre les k composantes principales donnerait

$$E(\rho_{ii}) = 1 \ (\forall i = 1, 2, ..., p).$$

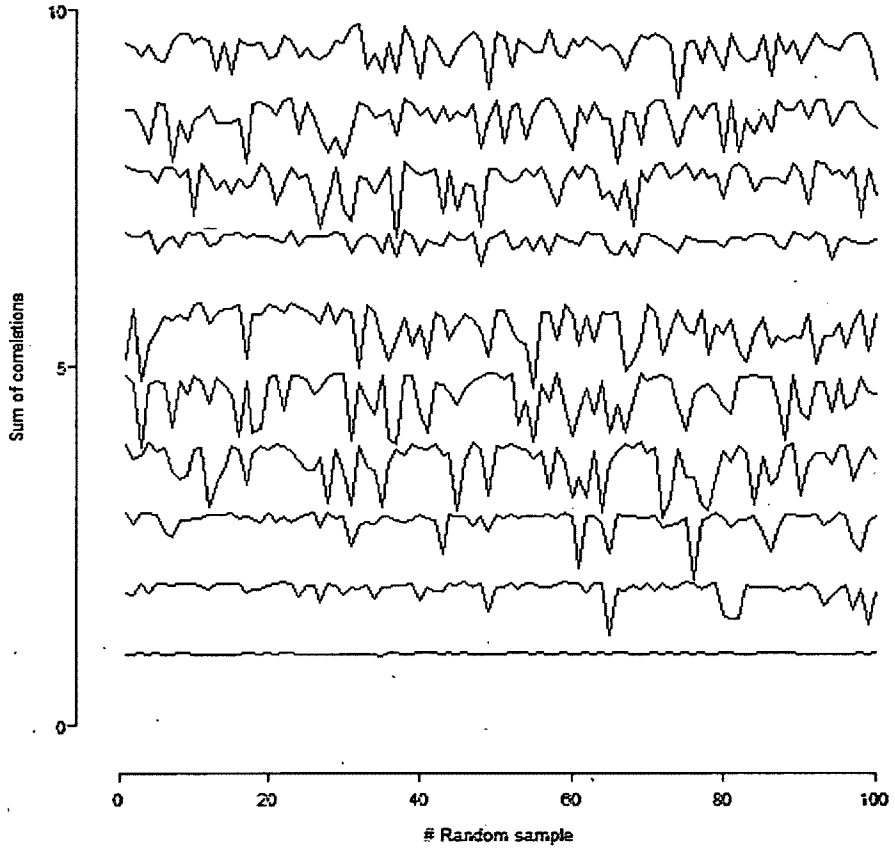
La mesure n'est pas sans faille cependant. En effet, il peut arriver que même si E_k est suffisamment stable, les k premières composantes principales ne le soient pas. C'est notamment le cas lorsque deux valeurs propres sont très près l'une de l'autre et engendrent une permutation lors du rééchantillonnage.

Pour corriger la situation, on utilisera le cas échéant la statistique suivante pour mesurer la stabilité globale des axes.

$$B_k = \sum_{i \leq k} \rho_{ii}$$

Enfin, on évalue la stabilité de la structure graphiquement en traçant chaque valeur de A_k ou B_k dans un graphique semblable à la figure 1.3.

Figure 2.3: Illustration des résultats de l'algorithme de Daudin,, Duby et Trecourt



On remarque que la première composante est extrêmement stable tandis que les composantes au-delà de la troisième le sont beaucoup moins.

2.4 Les scores alternatifs

Une dernière mesure de stabilité consiste à construire ce qu'on appelle les *scores alternatifs*. Les *scores alternatifs* sont en fait des scores calculés à partir des composantes principales d'une sous-population et des données de la matrice d'observations $X_{n\times p}$ de l'ensemble de la population.

En d'autres mots, on effectue le changement de repères de la matrice des observations $X_{n\times p}$ à partir des composantes principales $\alpha_i'^{(SP)}$ de la sous-population au lieu des composantes principales α_i' de la population. Cela nous permet de trouver

$$U^a_{n\times p} = X_{n\times p} \times A^{(SP)}_{p\times p} = \begin{pmatrix} u^a_{11} & u^a_{12} & \dots & \dots & u^a_{1p} \\ u^a_{21} & u^a_{22} & \dots & \dots & u^a_{2p} \\ . & . & & & . \\ . & & . & . & . \\ . & & & . & . \\ u^a_{n1} & u^a_{n2} & \dots & \dots & u^a_{np} \end{pmatrix}$$

où u^a_{ij} devient le résultat du produit scalaire de la j^e composante principale de la sous-population et de la i^e observation de la population.

Nous sommes ensuite en état de mesurer la stabilité de l'ACP en corrélant les *scores alternatifs* aux scores correspondant dans la population, c'est-à-dire en calculant

$$Corr(\mathbf{u}^a_i, \mathbf{u}_i) \ (\forall i = 1, 2, ..., p)$$

Une corrélation parfaite pour une certaine composante illustrera dans ce cas une stabilité parfaite de la composante concernée.

Chapitre 3

Le syndrome métabolique

Le concept de syndrome métabolique a été le sujet de plusieurs controverses cliniques et étiologiques ces dernières années. Pourtant, les associations entre les cinq facteurs de risque du syndrome métabolique (obésité, hypertension, hyperglycémie, taux de triglycérides élevé et taux de cholestérol HDL bas) nous permettent d'établir la validité du concept, spécialement dans une cohorte représentative d'une population réelle.

On utilise donc l'ACP pour analyser la structure des composantes physiologiques du syndrome métabolique sur 7213 patients enregistrés dans une banque de données administrative du CHUS, une population réelle avec plusieurs historiques médicaux différents afin de constater nous-mêmes s'il y a lieu d'avoir controverse. On valide ensuite les résultats en répétant les analyses sur plusieurs échantillons aléatoires ou non aléatoires à l'aide de procédés calqués sur le *bootstrap*. On utilise enfin le premier axe de l'ACP pour prédire certains types de maladie (diabète et maladies cardiovasculaires)

On trouve deux composantes principales expliquant 53% de la variance. La première composante (expliquant 33% de la variance) est associée, dans la direction prévue, avec les cinq variables prédictives nous amenant à l'interpréter comme le processus représentant le syndrome métabolique dans la structure de données.

Les résultats nous amènent à croire que le syndrome métabolique est bel et bien un processus existant.

De plus, la première composante est beaucoup plus prédictive des maladies subséquentes que les quatre autres composantes. Ainsi, un indicateur pourrait éventuellement être construit pour identifier le syndrome métabolique, et ce, même sans mesurer les cinq facteurs aggravants. On pourra donc mieux prévenir les risques de maladies telles que le diabète et les maladies cardiovasculaires à l'aide d'un outil clinique semblable qui pourrait être développé dans le futur.

En ce qui concerne l'analyse du syndrome métabolique, on considère ces caractéristiques :

1. La dimension de la masse de données est faible puisque le syndrome métabolique ne comprend que cinq facteurs. L'analyse est donc beaucoup plus facile à interpréter.
2. On connaît la structure théorique du syndrome. Ainsi, nous n'avons qu'à vérifier si la première composante correspond au syndrome métabolique. La validation du syndrome est donc relativement simple.
3. Il faut confirmer que les *charges* vont dans la bonne direction, c'est-à-dire vérifier si le cholestérol HDL varie dans le sens contraire des autres variables puisque le critère associé est un seuil maximal au lieu de minimal.
4. Il faut s'assurer que la structure est stable et non pas un artefact de notre échantillon.

La contribution de l'auteur (Francis Dusseault-Bélanger) correspond à 80% de la charge de travail reliée à la rédaction de l'article et à 100% de la charge de travail reliée aux analyses statistiques.

Validating metabolic syndrome through principal component analysis in a medically diverse, realistic cohort

Running title: Validating metabolic syndrome through PCA

Francis Dusseault-Belanger[1,4], BSc - Alan A Cohen[2,4,5], PhD - Marie-France Hivert[3,4,6], MD, MMSc - Josiane Courteau[4], PhD - Alain Vanasse[2,4], MD PhD

[1] Département de mathématiques, Université de Sherbrooke, Sherbrooke, Québec, Canada

[2] Département de médecine de famille, Université de Sherbrooke, Sherbrooke, Québec, Canada

[3] Département de médecine, Université de Sherbrooke, Sherbrooke, Québec, Canada

[4] Centre de Recherche Clinique Étienne-Lebel - Centre Hospitalier Universitaire de Sherbrooke, Sherbrooke, Québec, Canada

[5] Centre de recherche sur le vieillissement – Centre de santé et services sociaux – Institut universitaire de gériatrie de Sherbrooke, Sherbrooke, Québec, Canada

[6] Massachusetts General Hospital, Boston, Massachusetts, USA

Funding source: The Centre de Recherche Clinique Etienne-LeBel and the Centre de recherche sur le vieillissement are two clinical research centres supported by the Fonds de Recherche en Santé du Québec (FRSQ). Marie-France Hivert is supported by an FRSQ Scholar Award (junior 1 level) and a Canadian Diabetes Association Clinical Scientist Award. Alain Vanasse receives a career award form the FRSQ as a senior clinician scholar.

Corresponding author:

Alan Cohen

Groupe de recherche PRIMUS

3001, 12e avenue Nord, Sherbrooke

Québec, Canada, J1H 5N4

E-mail : alan.cohen@usherbrooke.ca

Phone : 819-346-1110 ext.12589

Abstract

Background: The concept of metabolic syndrome has been subject to etiological and clinical controversies in recent years. Associations among the five risk factors (obesity, hypertension, hyperglycemia, high triglyceride levels and low HDL cholesterol) may help establish the validity of the concept, especially in a cohort representative of an actual population.

Methods: We used principal component analysis (PCA) to analyze the structure of the physiological components of metabolic syndrome in 7213 patients contained in an administrative database for the CHUS hospital in Sherbrooke, Quebec, a realistic cohort with diverse medical histories. We validated the results by repeating the analysis on stratified and random subgroups of patients, and on different combinations of risk factors. The first axis of the PCA was used to predict coronary heart disease (CHD) and diabetes.

Results: The two first axes explained 53% of the variance. The first axis (33%) was associated in the expected direction with all five predictor variables, consistent with its interpretation as metabolic syndrome. The first axis was more predictive of subsequent CHD and diabetes than the formal definition of metabolic syndrome.

Conclusions: These results suggest that the concept of metabolic syndrome accurately captures an existing underlying physiological process. A continuous indicator could be constructed to identify more accurately metabolic syndrome thus improving risk assessment for CHD and diabetes mellitus. Metabolic syndrome can be measured well even without all five predictors. However, discrepancies with other studies suggest that our results may not be generalizable, perhaps because our cohort tends to be sicker.

Introduction

Over the last 20 years, the concept of metabolic syndrome has been gradually adopted to explain the physiological dysregulation underlying risk for diseases such as coronary heart disease (CHD) and diabetes mellitus. According to the National Cholesterol Education Program Adult Treatment Panel III (NCEP-ATPIII), metabolic syndrome is currently identified by the presence of at least three out of five risk factors: obesity, high blood pressure, high blood sugar, high triglyceride levels and low HDL cholesterol (NCEP-ATPIII[1] thresholds are given in Table 1). However, there is still some debate in the literature as to whether metabolic syndrome is indeed a real phenomenon based on one specific pathophysiogic aetiology or simply a collection of individual risk factors[2,3].

There have been several studies of the validity of metabolic syndrome using factor analysis but often disagree[4-10]. Some identifies a single syndrome[4] whereas others identify multiple independent processes related to metabolic syndrome[5-8]. Factor analysis is primarily used in the social sciences to evaluate questionnaires composed of items designed specifically to measure pre-identified constructs (latent variables). For this reason, factor analysis has come under some criticism as being overly subjective[11].

In contrast, principal component analysis (PCA) is an efficient representation of the correlation structure of a set of variables. As such, it is not useful for hypothesis testing but is the least biased way to understand patterns among variables and thereby to look at the underlying processes producing these patterns. In contrast to items on a questionnaire,

physiological variables are not devised in order to measure an *a priori* construct or process. Physiological data can thus be used to detect and measure such underlying processes. Thus, unless the processes are already well-identified and characterized, PCA is a preferable method because it imposes no biases or presuppositions other than the choice of variables[12]. PCA should thus allow us to see if something like metabolic syndrome emerges from the data when we are not designing models specifically to detect it.

The first question PCA can help us answer is: What is the dimensionality of the data? In the case of metabolic syndrome, dimensionality is a crucial question because the validity of the concept of a single syndrome hinges on our ability to describe it with a single variable. Each PCA axis is independent (orthogonal) from the others and can thus be considered a variable corresponding to a dimension, so if we require two axes to describe most of the variance in our data, it implies that at least two independent processes are present.

Second, how do the raw variables associate with each dimension? Again, this is a crucial question for metabolic syndrome. If the primary axis is tightly associated with all five predictors in the right direction, then we are right to interpret a single syndrome and we are right to include all five predictors in its definition. However, if only four predictors associate with this first axis, the missing predictor is not an important part of metabolic syndrome, or the concept as traditionally defined is not valid.

Finally, how does each individual score on each axis? At a clinical level, this should allow doctors to assess the composite risk for a single patient. At an analytical level, the ability to assign scores allows us to perform subsequent analyses including, for example,

an assessment of the predictive value of each axis for outcomes such as incidence of CHD and diabetes mellitus.

Our objectives are therefore to (1) validate the dimensionality of the syndrome; (2) propose a continuous indicator to define metabolic syndrome; (3) compare the new definition to the standard definition for their capacity to predict CHD and diabetes outcomes; and (4) assess whether metabolic syndrome can be well measured without all five predictors. We hypothesized that a clear PCA axis for metabolic syndrome would outperform the standard definition, which loses most of the original information by dichotomizing each of the five variables and then further dichotomizing the number of these present.

In addition, properly validating a PCA analysis allows further insight into the physiological processes determining the observed patterns. PCA is only valid when it is stable across subgroups of the population such as women or men, meaning that in each subgroup it will yield the same associations among the risk factors and thus be represented by an axis with the same interpretation (the metabolic syndrome). Conversely, a lack of stability could be informative: different axis structures for men and women would imply different regulatory networks and pathways of disease progression by sex. For metabolic syndrome, it is particularly important to validate that results are stable across sex, age, and disease status. The lack of any of these would tend to invalidate a general concept of metabolic syndrome and would suggest that pathways of disease progression and dysregulation are context dependent, a real possibility.

Methods

Data Source and Study Patients

The cohort studied is composed of patients older than 18 years old with at least one visit (outpatient or inpatient) at the Centre Hospitalier Universitaire de Sherbrooke (CHUS) from January 1st, 2002 to December 31, 2003. Vital signs, clinical data and anthropometric measures were collected from the patients along with data for International Classification of Diseases (ICD-9 and ICD-10 codes) and demographic information such as home postal code. We excluded patients having one or more missing data points on one of the 5 risk criteria for metabolic syndrome (weight as a clinical common measure of excess adiposity, arterial blood pressure, glycemia, HDL cholesterol, or triglycerides), leaving us with 7213 patients on the 71,151 patients available. A complete data set is necessary to conduct PCA without using any form of imputation methods. Data were obtained via the query system *Centre Informatisé de Recherche Évaluative en Service et Soins de Santé* (CIRESSS) which is directly linked to the clinical electronic health records of the institution; our cohort thus includes all patients meeting the above criteria.

Definition of Metabolic Syndrome

We compared three definitions of metabolic syndrome. (1) For the traditional definition, we used the updated NCEP-ATPIII thresholds to define formal criteria. The threshold for each of the metabolic syndrome's criteria are presented in Table 1. When formal criteria were missing, we looked for surrogate criteria based on results from a previous validation study[13]. When more than one measurement for a specific criterion

was available during the baseline period, we used the most extreme value (highest for all criteria except for HDL cholesterol where the lowest value was used). Patients were then categorized into two categories: No metabolic syndrome (less than three criteria met) and having metabolic syndrome (at least three criteria met). Frequencies of measurement and fulfillment of each criteria in the overall population were reported previously[14]. (2) We used the first PCA axis as a continuous score related to metabolic syndrome, a "degree of metabolic syndrome". (3) We also dichotomized the first PCA axis such that the patient either has metabolic syndrome (score greater than zero) or not (score lower than zero).To evaluate the contribution of blood pressure in the data with a single variable, we used mean arterial pressure (MAP), as is standard[15,16].

Outcome variables

Summary discharge forms provided the ICD codes used to identify the incidence of CHD (ICD-9 codes 410-414 and 428; ICD-10 codes I20-I25 and I50) and diabetes (ICD-9 code 250; ICD-10 codes E10-E14) without any discrimination between type 1 or type 2 diabetes. We collected those codes form January 1, 2000 through December 31, 2003 to assess prevalence of the diseases. We included only patients free of CHD and diabetes in order to assess the incidence of those conditions (this period length was chosen to ensure that we excluded 85% of prevalent diabetes cases, as validated previously[17]). The same codes were collected during a 5-year follow-up from January 1, 2004 to December 31, 2008 to identify incidence of CHD and diabetes. The use of ICD codes in administrative data to identify CHD has been previously validated[18,19].

Socio-demographic variables

Due to the fact that socioeconomic information was not available at the patient level, we used patients' postal codes to determine population quintiles of social and material deprivation using Pampalon and Raymond's index[20]. We also used patients' postal codes to assess whether the patient was living within or outside a Census Metropolitan Area (CMA). An area is considered a CMA if it has a population of at least 100,000, of which 50,000 or more live in the urban core[21], an approach validated in our population.

Statistical analysis

PCA was conducted on measures of weight, MAP, non-fasting glucose, HDL cholesterol and triglycerides to assess the dimensionality of the corresponding structure using Horn's parallel analysis[22]. The stability of each of the axes under random sampling was then tested on 100 random samples using bootstrap methods based on Daudin's algorithm[23]. We also modified the algorithm to assess the stability of the axes under nonrandom sampling using 14 subsamples regrouping patients into different categories. The categories used are as follows: (1) Married (2) Unmarried (3) Living in a CMA (4) Not living in a CMA (5) In the first (poorest) quintile of the material deprivation scale (6) In the second and third quintiles of material deprivation (7) In the fourth and fifth (richest) quintiles of material deprivation (8) In the first and second (poorest) quintiles of the social deprivation scale (9) In the third and fourth quintiles of social deprivation (10) In the fifth (richest) quintile of social deprivation (11) Men (12) Women (13) Under 65 years old (14) Over 65 years old. Repartition of the population between each subsample is shown in Table 2.

For each subsample, we extracted the loadings for all meaningful axes (based on percent variance explained) to calculate new alternative scores for the full population. We then compared their stability by assessing the pairwise correlations between these alternative scores and the scores from the complete dataset. The last stability assessment test computed correlations between the first axes given by the full PCA (Weight-MAP-Glucose-HDL-Triglycerides) to the ones emerging from a model including all possible subsets of the five criteria (such as Weight-MAP-HDL-Triglycerides).

To test the predictive value of metabolic syndrome, we used logistic regression to compute odd ratios for incidence of CHD and diabetes mellitus as a function of (1) MetS diagnosed by the standard NCEP-ATPIII definition (2) the PCA axes as continuous variables and (3) MetS as diagnosed by a dichotomous PCA criterion (greater vs. less than zero). The criterion was chosen so that we could compare two populations with approximately the same size. Analyses were adjusted for age, sex, and comorbidities. All analyses were performed using R version 2.14.1 (R Foundation for Statistical Computing, Vienna, Austria)

Results

There were 7213 patients available for PCA in the baseline period, including 4173 patients with CHD and/or diabetes. In this population, 4074 patients had metabolic syndrome by the standard definition (Table 2). We conducted PCA on the data to assess the dimensionality of the five metabolic syndrome variables system (Table 3). The first axis explains 33% of the variance, better than any of the original criteria (which would

each explain 20% of the variance if independent), the second performs about as well as any criterion, and the remaining three perform substantially worse. Horn's parallel analysis[22] identifies the first axis as being essential (adjusted eigenvalue of 1.625), but is inconclusive regarding the second axis (adjusted eigenvalue of 0.998). The analysis further eliminates the other three axes (adjusted eigenvalues lower than 0.854), meaning that we could capture most of the information of the metabolic syndrome with only two components. We can therefore consider the dimensionality of the five risk criteria to be at most two axes. The variable loadings (Table 3) confirmed that the first axis was consistent with metabolic syndrome: it was associated in the proper direction with all five predictor variables (positively with all factors except HDL cholesterol, whose threshold is an upper bound instead of a lower bound). The loadings for the remaining axes did not seem to correspond to known biological processes, suggesting that they are not related to metabolic syndrome. Use of only systolic or diastolic blood pressure instead of MAP did not substantively change the results (data not shown).

The stability of the axis structure was first verified with one hundred random bootstrap samplings using Daudin's algorithm[23]. The lack of variation in the first two axes suggests that they are stable (Figure 1), whereas the third axis is highly unstable, indicating that the third, fourth and fifth axes represent mainly noise and/or measurement error. The same result was found for the fourteen subgroup analyses (Figure 2).

The stability under different subsets of variables was also tested. The first axis' scores from the five-criteria PCA correlated well with those emerging from PCA using only three or four criteria (all correlations greater than 0.75), but correlations were

weaker for the one- or two- criteria models (Figure 3). Therefore, the first axis could be measured well even without all five predictors.

For each subsample, we extracted the loadings for all meaningful axes to calculate new alternative scores for the full population. The mean Pearson correlation coefficient obtained from this analysis was greater than 0.99 for the two first axes. In particular, the scores generated from the male and female subsamples, as well as the under 65 and over 65 years old subsamples, were particularly strongly correlated (all four over 0.997).

Finally, the ability to predict CHD and diabetes outcomes was compared between the standard definition, the dichotomized first PCA axis (with or without adjustment for the second PCA axis), and the fully continuous first PCA axis (Table 4). Comparison of odds ratios shows that the dichotomized first PCA axis always outperforms the standard definition, and that this effect is particularly strong for diabetes. Using the pseudo R-squared, we see that the continuous measure always provides a modest improvement over the dichotomous measure (Table 4). Inclusion of the second PCA axis had no noticeable effect on the models.

Discussion

This study examined the correlation structure of metabolic syndrome using PCA on data from the electronic health records of 7213 patients at the CHUS hospital. This correlation structure identified a process (measured as a PCA axis) associated with all five criteria of the metabolic syndrome and consistent with previous descriptions of the syndrome. HDL cholesterol, weight, triglycerides and glucose contributed strongly to this

axis, whereas blood pressure was associated more weakly. As seen in Figure 1, blood pressure contributed more to a second independent axis of the PCA, suggesting that it may be related to metabolic syndrome only secondarily to other mechanisms. These findings, especially the small contribution of the blood pressure, are consistent with previous studies[9,10]. But our study also differs from others by identifying only two relevant axes compared to often three or more axes in other studies[5-10]. This could be due to the fact that only five criteria were included in our analysis instead of the common ten or more criteria (often somewhat redundant) in other papers[5-7]. But even taking the number of axes into consideration, Lindblad et al.[8] and Meigs et al.[10] do identify a first PCA axis similar to the one we found.

Subgroup analyses in men and women gave highly consistent results, meaning that the syndrome is not sex-specific in our population. Similarly, no differences in the syndrome were identified in younger (< 65 years old) or older subgroups, or in other population subgroups. Taken together, these analyses suggest that metabolic syndrome is a stable and highly reproducible phenomenon, at least in the Quebec context.

Lastly, we were able to demonstrate that subsets of MetS criteria provide nearly as good a measure of metabolic syndrome as all five together. When various PCA models including four or less criteria were correlated with the full model, all models including at least 3 criteria produced an axis strongly correlated with that from the full model. This suggests that many clinicians could identify metabolic syndrome without recommending further exams based solely on information already contained in electronic records. Additionally, hospital-wide screening may help identify at-risk patients, allowing clinicians to intervene more rapidly to prevent CHD and diabetes.

Limitations

(1) Our data come from hospital records, largely in-patient, meaning that we are studying a substantially sicker population than the general population. (2) The sample was restricted to patients having all five criteria measured. Accordingly, patients had many visits to the hospital and took many tests, further suggesting that they represent a sick population. This agrees with our higher-than-usual metabolic syndrome prevalence of 56%, compared to a prevalence of ~25% in most populations[24]. This could explain differences with other studies on healthier populations[5,7,8]. (3) The lack of information about patient treatments could have resulted in patients being misclassified with respect to specific criteria. (4) The overall population is 97% North American Caucasian, perhaps explaining some discrepancies with other studies[5,7,9].

Conclusion

In summary, we detected two processes determining the five variables related to metabolic syndrome in a sick Eastern Townships population. We clearly identify the most important process, the first PCA axis, as related to metabolic syndrome, whereas the second process is not related to it (strongly associated with arterial blood pressure). Furthermore, the first axis strongly predicts risk of diabetes or CHD incidence, whereas the second axis adds no additional predictive value.

Therefore, our results suggest the possibility for building an improved clinical measure of metabolic syndrome based on PCA that would provide a more precise

indication of CHD or diabetes risk. We built such a tool by using the linear combination of the loadings identified as the first PCA axis, dichotomized at zero, to test the efficiency of the tool to predict the discussed diseases. We find that the tool predicts diabetes much better than the standard definition, and is a slight improvement for CHD. Previous studies have also shown that metabolic syndrome is more strongly associated with diabetes incidence than with CHD events[25,26].

This improvement in predictive value likely comes from the fact that the PCA allows us to use continuous measures to summarize the information instead of five dichotomous measures like the metabolic syndrome. This can be seen in other studies as well[27]. A completely continuous tool would offer slightly more precision but may be difficult to implant in clinical practice.

Disclosure

The authors declare that no competing financial interests exist.

References

[1] Expert Panel on Detection, Evaluation and Treatment of High Blood Cholesterol in Adults. *Executive summary of the third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation and treatment of high blood cholesterol in adults (Adult Treatment Panel III).* See comment. JAMA, 2001, 285:2486-97.

[2] Kahn R, Buse J, Ferrannini E, et al. *The Metabolic Syndrome: Time for a Critical Appraisal.* Diabetes Care, 2005, 28:2289-2304.

[3] Neel JV, Julius S, Weder A, et al. *Syndrome X: Is it for Real ?* Genet Epidemiol, 1998, 15:19-32.

[4] Shen BJ, Todaro JF, Niaura R, et al. *Are Metabolic Risk Factors One Unified Syndrome? Modeling the Structure of the Metabolic Syndrome X.* Am J Epidemiol, 2003, 157:701-711.

[5] Oh JY, Sung YA, Hong YS, et al. *Prevalence and factor Analysis of Metabolic Syndrome in an Urban Korean Population.* Diabetes care, 2004, 27:2027-2032.

[6] Edwards, KL, Austin MA, Newman B, et al. *Multivariate Analysis of the Insulin resistance Syndrome in Women. Arterioscler Thromb Vasc Biol,* 1994, 14:1940-1945.

[7] Hanson RL, Imperatore G, Bennett PH, et al. *Components of the "Metabolic Syndrome" and Incidence of Type 2 Diabetes.* Diabetes, 2002, 51:3120-3127.

[8] Gray RS, Fabsitz RR, Cowan LD, et al. *Risk Factor Clustering in the Insulin Resistance Syndrome. The Strong Heart Study.* Am J Epidemiol, 1998, 148:869-878.

[9] Lindblad U, Langer RD, Wingard DL, et al. *Metabolic Syndrome and Ischemic Disease in Elderly Men and Women.* Am J Epidemiol, 2001, 153:481-489.

[10] Meigs JB, D'Agostino RB, Wilson PWF, et al. *Risk Variable Clustering in the Insulin Resistance Syndrome: The Framingham Offspring Study.* Diabetes, 1997, 46:1594-1600.

[11] Lawlor DA, Ebrahim S, May M, et al. *(Mis)use of Factor Analysis in the Study of Insulin resistance Syndrome.* Am J Epidemiol, 2004, 159:1013-1018.

[12] Velicer WF, Jackson DN. *Component Analysis versus Common Factor Analysis: Somme Issues in Selecting an Appropriate Procedure.* Multivariate Behav Res, 1990, 25:1-28.

[13] Hivert MF, Grant R, Shrader P, et al. *Identifying Primary Care Patients at Risk for Future Diabetes and cardiovascular Disease Using Electronic Health records.* BMC Health Serv Res, 2009, 9:170.

[14] Hivert MF, Dusseault-Bélanger F, Cohen A, et al. *Modified Metabolic Syndrome Criteria for Identification of Patients at Risk of Developing Diabetes and Coronary Heart Diseases: Longitudinal Assessment via Electronic Health records.* Can J Cardiol, 2012. In press.

[15] Safar ME, Frohlich ED. *The Arterial System in Hypertension: A Prospective View.* Hypertension, 1995, 26:10-14.

[16] Henry RM, Kostense PJ, Spijkerman AM, et al. *Arterial stiffness increases with deteriorating glucose tolerance status: the Hoorn Study.* Circulation, 2003, 107:2089-2095.

[17] Asghari S, Courteau J, Carpentier AC, et al. *Optimal Strategy to Identify Incidence of Diagnostic of Diabetes Using Administrative Data.* BMC Med Res Methodol, 2009, 9:62.

[18] Levy AR, Tamblyn RM, Fitchett D, et al. *Coding accuracy of hospital discharge data for elderly survivors of myocardial infarction.* Can J Cardiol, 1999, 15:1277-1282.

[19] Petersen LA, Wright S, Normand SL, et al. *Positive predictive value of the diagnosis of acute myocardial infarction in an administrative database.* J Gen Intern Med, 1999, 14:555-558.

[20] Pampalon R, Raymond G. *A deprivation index for health and welfare planning in Quebec.* Chronic Dis Can, 2000, 21:104-113.

[21] Statistics Canada. Consulted August 2010.
http://www12.statcan.ca/english/census01/products/reference/dict/geo009.htm

[22] Horn JL. *A Rationale and test for the Number of Factors in Factor Analysis.* Psychometrika, 1965, 30:179-185.

[23] Daudin JJ, Duby C, Trecourt P. *Stability of Principal Component Analysis Studied by the Bootstrap Method.* Statistics, 1988, 13:405-410.

[24] Cameron AJ, Shaw JE, Zimmet PZ. *The metabolic syndrome: prevalence in worldwide populations.* Endocrinol Metab Clin North Am, 2004, 33:351-375.

[25] Sattar N, McConnachie A, Shaper AG et al. *Can metabolic syndrome usefully predict cardiovascular disease and diabetes? outcome data from two prospective studies.* Lancet, 2008, 371:1927-1935.

[26] Wilson PW, D'Agostino RB, Parise H, et al. *Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus.* Circulation, 2005, 112:3066-3072.

[27] Wilson PW, Meigs JB, Sullivan L, et al. *Prediction of Incident Diabetes Mellitus in Middle-aged Adult: The Framingham Offspring study.* Arch Intern Med, 2007, 167:1068-1074.

Chapitre 4

Le processus du vieillissement

Alors que plusieurs études ont tenté d'identifier les associations existant entre différents biomarqueurs et le vieillissement, peu d'études ont tenté d'identifier ces associations en intégrant l'information de plusieurs biomarqueurs simultanément. Cela peut être dû au défi méthodologique entourant une telle recherche étant donné le nombre impressionnant de biomarqueurs humains et nos connaissances superficielles concernant le vieillissement.

Pour relever ce défi, on utilise l'ACP afin d'explorer la structure de corrélation et sa relation avec l'âge à travers 55 biomarqueurs sanguins mesurés sur 1226 femmes en santé âgées de 65 à 102 ans résidant près de Baltimore (Maryland) aux États-Unis sur une période de 15 ans.

Premièrement, on identifie les axes primaires de variation des données ainsi que leur interprétation dans le contexte basée sur les *charges* de chacune des composantes principales retenues. On trouve 18 axes pertinents qui expliquent ensemble 66% de la variance des données, soit suffisamment pour potentiellement être importants.

Deuxièmement, on utilise des méthodes de *bootstrap* pour évaluer la stabilité de la structure décelée. La première composante principale est très stable tandis que les deuxième et troisième composantes le sont modérément. Quant aux dernières composantes, elles sont toutes très instables suggérant une certaine dépendance de la structure de corrélation au contexte de l'étude.

En ce qui a trait à l'interprétation, les trois premiers axes expliquent 21% de la variance et les biomarqueurs leurs étant associés sont un amalgame de plusieurs systèmes physiologiques différents ne suggérant aucun lien avec des théories connues. De plus, seulement la deuxième composante semble avoir une corrélation avec l'âge, bien que très petite.

Malgré l'association minimale avec l'âge et l'interprétation biologique floue, la stabilité des trois premières composantes nous laisse croire qu'il pourrait exister un processus physiologique régissant les activités des différents biomarqueurs étudiés même s'il n'est pas un processus clé du vieillissement.

En ce qui a trait à la méthodologie, le principe de validation sera beaucoup plus complexe puisqu'on connaît mal la structure étudiée :

1. La dimension de la structure de données est élevée. L'analyse est donc beaucoup plus difficile à interpréter.
2. On connaît mal la structure théorique du processus. Déceler un processus régissant le système de biomarqueurs sera donc plus complexe puisque nous ne savons pas ce qu'on cherche.
3. On doit s'assurer que la structure est stable et non pas un artefact de l'échantillon.

La contribution de l'auteur (Francis Dusseault-Bélanger) correspond à 80% de la charge de travail reliée à la rédaction de l'article et à 100% de la charge de travail reliée aux analyses statistiques.

Variance in 55 clinical blood biomarkers is largely unassociated with age in the Women's Health and Aging Study

Running title:

Francis Dusseault-Bélanger[1,3], BSc – Alan A Cohen[2-4], PhD – Qian-Li Xue[5], PhD – Linda P Fried[6], MD, MPH

[1] Département de mathématiques, Université de Sherbrooke, Sherbrooke, Québec, Canada
[2] Département de médecine de famille, Université de Sherbrooke, Sherbrooke, Québec, Canada
[3] Centre de Recherche Clinique Étienne-Le Bel - Centre Hospitalier Universitaire de Sherbrooke, Sherbrooke, Québec, Canada
[4] Centre de recherche sur le vieillissement – Centre de santé et services sociaux – Institut universitaire de gériatrie de Sherbrooke, Sherbrooke, Québec, Canada
[5] Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA
[6] Columbia University Mailman School of Public Health, New York, New York, USA

Funding source: The Centre de Recherche Clinique Étienne-Le Bel and the Centre de recherche sur le vieillissement are two clinical research centres supported by the Fonds de Recherche en Santé du Québec (FRSQ).

Corresponding author:
Alan Cohen
Groupe de recherche PRIMUS
3001, 12e avenue Nord, Sherbrooke
Québec, Canada, J1H 5N4

E-mail : alan.cohen@usherbrooke.ca
Phone : 819-346-1110 ext.12589

Running page headline: Biomarkers variance not related to age

Abstract

While many studies have assessed associations between individual biomarkers and aging, few studies have attempted to identify physiological processes in aging based on integrating information from multiple markers. We used principal component analysis (PCA) to explain correlation structure and age associations underlying 55 standard clinical blood biomarkers measured in the Women's Health and Aging Study. The first 18 axes explained a total of 66% of the variance, enough to potentially be important. The three first axes are stable (explaining 21% of the variance), and the remaining axes are highly unstable, suggesting condition-dependence of the correlation structure. Only the second axis shows a more important (if weak) correlation with age ($r=0.14$, $p<0.0001$). Despite the minimal associations with age and the unclear biological interpretation, the stability of the first three axes suggests they may represent hither-to unknown physiological processes driving associations among biomarkers, though they do not appear to be key to aging.

Introduction

Many disciplines have been involved in the search for underlying causes of aging, including biology, epidemiology, and demography (1-3). In recent years, the search for a single cause of aging is starting to give way to a more complex view in which aging is a multifactorial process (4,5). In particular, the aging process may be due wholly or in part to a systematic dysregulation of physiological homeostatic processes (6-8). Measuring such dysregulation is a substantial challenge, given that the regulatory structure of the underlying network is largely unknown; nonetheless, it is likely that underlying dysregulatory processes would leave a signal that could be detected through multivariate statistical analysis of physiological and/or biochemical profiles such as gene expression data, metabolomics, or peripheral blood biomarkers.

Here, we hypothesized that a large suite of biomarkers – even biomarkers not known to be associated with aging – would contain a strong signal of underlying physiological processes, and that these processes might be associated with aging. Many statistical approaches could be applied to these questions; here we use one of the most straightforward, principal component analysis (PCA). In contrast to the few previous studies using PCA on aging biomarkers, we were not attempting to quantify biological age (9-11) but rather to understand processes that may become dysregulated during aging. Thus, we were possibly searching for multiple axes, not for a single axis, and were not constrained to biomarkers with known correlations with age (12). By definition, PCA is the most efficient representation of the correlation structure of a set of variables. PCA should thus allow us to see if patterns emerge from the data when we are not designing models specifically to detect it based on *a priori* biological hypotheses. Such exploratory approaches are justified given our poor comprehension of the underlying system, though hypothesis-driven approaches may also be fruitful.

PCA works by creating a set of new variables ("axes") equal in number to the original set that are linear combinations of the original set. The combinations are chosen so as to maximize the variance explained by the first axis, maximize the remaining variance explained by the second axis, and so forth,

on the condition that each axis be independent of previous axes. The importance of each axis is thus identified by the variance it explains, and the interpretation of each axis is based on its associations ("loadings") with the original variables. The loadings can be used to calculate "scores" for the patients based on the linear combinations of their observed biomarker levels, so that the new axes serve as new summary variables. At an analytical level, the ability to assign scores allows us to perform subsequent analyses including, for example, an assessment of the predictive value of each axis for disease outcomes or its correlation with age. The importance of the observed correlation structure should be validated by assessing its stability across population subgroups (13-14). Properly validating a PCA analysis could allow further insight into the physiological processes determining the observed patterns. It is possible that PCA is only valid when it is stable across subgroups of the population, meaning that in each subgroup it will yield similar associations among biomarkers and thus be represented by axes with the same interpretations. Conversely, a lack of stability could be informative: different axis structures for Caucasians and African-Americans might imply different regulatory pathways by race for example. For our study, it is particularly important to validate that results are stable across different subpopulations. A lack of stability would suggest that pathways of dysregulation are context dependent, a real possibility.

Here, we used PCA on 55 standard clinical biomarkers taken on 818 participants aged 65 to 102 years old from the Women's Health and Aging Study (WHAS). Using multiple validation measures on both static markers and their changes over time, we assessed the importance and stability of the primary axes, as well as their interpretation based on their loadings. Finally, we assessed whether the axes correlated with age.

Methods

Data Source and Study Patients

The Women's Health and Aging Study is a population-based prospective study of community-

dwelling women drawn from Baltimore City and Baltimore County, Maryland between 1993 and 2008. Specifically, our analyses used merged data from WHAS I () and WHAS II (). WHAS I was composed of 1002 participants drawn from the one-third most disabled portion of the population of women aged 65+, whereas WHAS II studied 436 women among the two-thirds least disabled women aged 70 - 79 years old at baseline. WHAS I required a Mini-Mental State Examination () score above 17 and WHAS II a score above 23. All evaluations were conducted using the same standardized methods. Subjects were followed up 1.5, 3, 6, 7.5 and 9 years later. We used 55 different biomarkers collected through the study including blood basics, immune measures, hormones, inflammatory measures, ions, lipids, oxygen transport measures, protein metabolism, sugar metabolism, micronutrients and vitamins (Details shown in supplementary table). We excluded any visits having one or more missing values on any of these 55 variables, leaving us with 1230 visits from 818 patients. Individual patients had up to three visits. We also studied the changes in biomarkers between rounds 1 and 2 (n=187), calculated simply as the difference between the values.

Statistical analysis

PCA was conducted on the 55 biomarkers and potentially important axes were identified using Horn's parallel analysis (15). The stability of each of the axes under random sampling was then tested on 100 random samples using bootstrap methods based on Daudin's algorithm (16). We also modified the algorithm to assess the stability of the axes under nonrandom sampling using 13 subsamples regrouping patients into different categories instead of a hundred random samples. The subsamples were chosen based on available demographic variables thought to potentially be linked to important physiological differences. The different categories used are as follows: (1) WHAS1 (2) WHAS2 (3) White (4) Black (5) Between 70 and 80 years old (6) Over 80 years old (7) Married (8) Not married (9) Less than 10 years of education(10) More than 10 years of education (11) < 10,000$ (12) Annual income between 10,000$ and 20,000$ (13) Annual income > 20,000$. Repartition of the population

between subsamples is shown in Table 1. Stacked bar plots of the associations of the original variables with the axes were used to study the stability of the composition of the first three axes across subsamples. The same thing was done to compare the first three axes from all three data subsets (all visits, first visit of all patients, and all differences between round 1 and round 2). Comparisons of loading structure across alternative datasets were based on the 20 variables explaining the most variance in the relevant axis calculated for the full data set (i.e., 1230 visits).

For each subsample, we extracted the loadings for the first axis to calculate new alternative scores for the full population. We built these alternative scores by combining the loadings from the subsample with the data from the entire population. We then compared their stability by assessing the pairwise correlations between these alternative scores and the real scores from the complete data. These analyses were repeated using only the first visit for each patient instead of a maximum of three visits per patients and by using the difference in biomarkers between round 1 and round 2. We compared the results to assess the stability of the structure's composition. Finally, we assessed the correlations between the axes and age. All analyses were performed using R version 2.14.1 (R Foundation for Statistical Computing, Vienna, Austria)

Results

Results of the primary analysis on all 1230 visits showed that the first 19 axes explain between 1.9% and 8.7% of the total variance each, more variance than any of the original biomarkers (1/55 =1.8%). Horn's parallel analysis (15) identifies the first 17 axes as being the most significant in the correlation structure (all adjusted eigenvalues greater than 1). Thus there are potentially as many as 17 to 19 axes or processes defining the system.

The stability of these axes was then verified with one hundred random bootstrap samplings using Daudin's algorithm (16) (Figure 1). The lack of variance on the first axis suggests that it is quite stable, whereas the second and third axes are moderately stable and the remaining axes are clearly unstable,

suggesting high levels of context-dependence in much of the structure. Similar results were found with the fourteen subgroup analyses (Figure 2).

For each subsample, we extracted the loadings for the first axis to calculate the new alternative scores for the entire population. We then compared these alternatives scores with the original scores using Pearson r. The mean Pearson r obtained from this analysis is greater than 0.99 for the first axis whereas the mean drops to 0.76 and 0.71 for the second and third axes respectively.

Finally, we compared the composition of the first three axes and one more distant axis across the 13 subsamples. As seen in Figure 3, the axis composition is quite stable for the first axis despite some subsamples containing less than 200 visits. The variables associated with the first axis come from a variety of different systems, making its biological interpretation difficult (Supplementary table and Figure 3). The second and third axes are unstable in several subgroups but are consistent in most (Figure 3). For comparison, we also show the 12the axis, which is highly unstable across subgroups.

The first axis was not correlated with age (r=-0.0038, p=0,79). The second axis was weakly correlated (r=0.14, p<0.0001), and all remaining axes were uncorrelated (Table 2). Analyses were repeated using only one visit per patient and using changes in biomarkers between round 1 and round 2, but few differences were observed in variance explained (Table 3) or in axis composition despite major differences in dataset size.

Discussion

This study examined whether the correlation structure among 55 standard clinical biomarkers was informative about the aging process in a population of 818 elderly women between 1993 and 2008 in Baltimore, Maryland through the Women Health and Aging Study (WHAS). We did not find strong evidence for an association between the major axes of correlation structure and age, nor did we find a difference when we looked at changes in biomarkers levels over time instead of static levels. However, the first axis was highly stable, indicating that it may represent an important new physiological process

unrelated to aging.

The absence of a relationship with age could be explained by multiple factors: (1) There may be very little signal of aging among the 55 biomarkers used; (2) homogeneity of ages of participants (mostly 70-85) may make it hard to detect signals that might have been apparent had we included more young or very old women; and (3) the aging process might differ substantially from one individual to another, making it impossible to identify through PCA.

The stability of the first PCA axis suggests it may represent a real physiological process, but its composition does not appear to relate directly to any known biological process. For example, the 6 most important contributors to the axis come from very different biological groupings such as transport proteins (albumin, albumin/globulin ratio) and oxygen delivery (hemoglobin, hematocrit, iron, saturated iron). Potentially, the first axis is related to hepcidine (17), since this protein is related to iron metabolism and inflammation (18). We expected to find a few reliable axes representing one or two biological groupings of biomarkers each instead of such an amalgamated axis. The stability of the axis is even more interesting considering its unusual composition and its lack of correlation with age. Subgroup analyses gave highly consistent results, meaning that the process is a stable and highly reproducible phenomenon, at least in the WHAS context (community-dwelling elderly women in Baltimore).

The second and third axes have more ambiguous stability, though the second was significantly (if weakly) correlated with age. The second axis is mainly composed of sugar metabolism markers (protein-bound glucose, glycated hemoglobin and glucose) and lipids (cholesterol, HDL cholesterol and triglycerides), suggesting a relationship to metabolic syndrome (13). The third axis is composed of basic blood markers (red blood cells and mean corpuscular hemoglobin) and circulating proteins (direct bilirubin, alanine and aspartate transaminases). Further studies will be required to assess whether these axes can be replicated in other populations and whether they represent underlying biological processes or unstable patterns caused by population heterogeneity.

Other authors have used PCA to estimate biological age. Nakamura et al. (9) found that five variables (out of 29) explained the most out of biological variance: forced expiratory volume in 1 second, systolic blood pressure, hematocrit, albumin, and blood urea nitrogen. Other studies differ from ours by pre-selecting biomarkers which are already correlated with age (10,11). Ueno et al. (10) found that forced expiratory volume in 1 second, systolic blood pressure, mean corpuscular hemoglobin, glucose and albumin/globulin ratio were the most relevent to estimate biological variance from a set of 31 variables. Park et al. (11) developped an estimation of biological age using VO(2)max, percent body fat, waist circumference, forced expiratory volume in 1 second, systolic blood pressure, low density cholesterol, blood urea nitrogen, serum albumin, erythrocyte sedimentation rate, hearing threshold and glycosylated hemoglobin. None of these results are inconsistent with what we present, though the different goals and biomarkers used make direct comparison difficult. Notably, Nakamura et al. (9) and Ueno et al. (10) also identified several very basic blood markers as important predictors.

Broadly speaking, PCA appears to be unable to identify key underlying aging processes, at least among the biomarkers studied in this dataset. Whether it has identified an important physiological process unrelated to aging but related to haematopoiesis, oxygen delivery, and protein transport remains to be seen. Regardless, our results are consistent with a substantial complexity in the correlation structure of blood biomarkers during aging, as should be expected for a complex dynamic system. Future studies should explore other promising methods such as multivariate statistical distance or structural equation models (SEM). It may then be possible to develop multivariate biomarkers of aging (19) to predict the rate of aging or monitor basic processes underlying disease states.

Limitations

(1) The conclusions and data illustrated come from a dataset containing more than one visit per patient. In theory, this could lead to dependence problems in our data, though no major differences were observed when analyses were repeated using only one visit per individual. (2) Our population

contains exclusively elderly women. A different data set would have to be used to generalize our results to men, or across the life course. (3) Our relatively homogeneous population could result in a lack of variation among biomarker levels and profiles, reducing our capacity to detect important patterns.

Fundings

This work was supported by the Canadian Institute of Health Research (grant numbers 110789, 120305, 119485); and the Natural Sciences ans Engineering Research Council (grant number 402079-2011).

Acknowledgements

AAC is a member of the FRQS-supported *Centre de recherché sur le vieillissement* and *Centre de recherche Étienne Le-Bel,* and is a funded Research Scholar of the FRQS.

References

[1] Gavrilov LA, Gavrilova NS. *The Reliability Theory of Aging and Longetivity.* J Theor Biol, 2001, 213:527-545.

[2] Puca AA, Daly MJ, Brester SJ et al. *A genome-wide scan for linkage to human exceptional longevity identifies a locus on chromosome 4.* Proc Natl Acad Sci USA, 2001, 98:10505-10508.

[3] Preston SH, Himes C, Eggers M. *Demographic Conditions Responsible for Population Aging.* Demography, 1989, 26:691-704.

[4] Weinert BT, and Timiras PS. *Invited Review: Theories of aging.* J Appl Physiol, 2003, 95:1706-1716.

[5] Kowald A, Kirkwood TB. *A network theory of aging: the intercations of defective mithocondria, aberrant proteins, free radicals and scavengers in the ageing process.* Mutat Res, 1996, 316:209-236..

[6] Varadhan R, Seplaki CL, Xue QL, Bandeen-Roche K, Fried LP. *Stimulus-response paradigm for*

characterizing the loss of resilience in homeostatic regulation associated with frailty. Mech Ageing Dev, 2008, 129:666-670.

[7] Seplaki CL, Goldman N, Glei D, Weinstein M. *A comparative analysis of measurement approaches for physiological dysregulation in an older population*. Exp Gerontol, 2005, 40:438-449.

[8] McEwen BS. *Stress, adaptation, and disease. Allostasis and allostatic load*. Ann NY Acad Sci, 1998, 840:33-44.

[9] Nakamura E, Miyao K, Ozeki T. *Assessment of biological age by principal component analysis*. Mech Ageing Dev, 1988, 46:1-18.

[10] Ueno LM, Yamashita Y, Moritani T, Nakamura E. *Biomarkers of aging in women ans the rate of longitudinal changes*. J Physiol Anthropol Appl Human Sci, 2003, 22:37-46.

[11] Park J, Cho B, Kwon H, Lee C. *Developing a biological age assessment equation using principal component analysis and clinical biomarkers of aging in Korean men*. Arch Gerontol Geriatr, 2009, 49:7-12.

[12] Crimmins EM, Johnston M, Hayward M, Seeman T. *Age differences in allostatic load: an index of physiological dysregulation*. Exp Gerontol, 2003, 38:731-734.

[13] Dusseault-Bélanger F, Cohen AA, Hivert MF, Courteau J, Vanasse A. Validating *Metabolic Syndrome Through Principal Component Analysis in a Medically Diverse, Realistic Cohort*, Metab Syndr Relat Disord, 2012. In press.

[14] Cohen AA, Dhingra N, Jotkat RM, Rodriguez PS, Sharma VP, Jha P. *Athe Summary Index of Malaria Surveillance (SIMS): a stable index of malaria within India, Popul Health Metr, 2010, 8:1.*

[15] Horn JL. *A Rationale and test for the Number of Factors in Factor Analysis.* Psychometrika, 1965, 30:179-185.

[16] Daudin JJ, Duby C, Trecourt P. *Stability of Principal Component Analysis Studied by the Bootstrap Method. Statistics, 1988, 13:405-410.*

[17] Ganz T. *Hepcidin, a key regulator of iron metabolism and mediator of anemia of inflammation.*

Blood, 2003, 102:783-788.

[18] Nemeth E, Ganz T. *Regulation of Iron Metabolism by Hepcidin.* Annu Rev Nutr, 2006, 26:323-342.

[19] Baker GT, Sprott RL. *Biomarkers of aging*. Exp Gerontol, 1988, 23:223-239.

Table 1: Characteristics of the women aged 65 to 102 years old in Baltimore included in the WHAS study between 1993 and 2008.

	Patients (first visit only)	Visits	Difference between round 1 and round 2
Total patients	818	1230	187
WHAS1	455 (56)	637 (52)	64 (34)
WHAS2	363 (44)	593 (48)	123 (56)
Caucasian	622 (76)	946 (77)	149 (80)
Married	256 (31)	389 (32)	55 (29)
Age	77.37 ± 6.14	77.20 ± 6.40	78.86 ± 5.41
Less than 70 years old	64 (8)	80 (6)	11 (6)
Between 70 and 80 years old	512 (63)	836 (68)	138 (74)
Over 80 years old	242 (29)	314 (26)	26 (14)
Education			
Less than 10 years	334 (41)	481 (39)	63 (34)
More than 10 years	480 (59)	744 (60)	124 (66)
Annual income	20,068 ± 22,143	20,746 ± 23,415	23,980 ± 20,582
Lower than $10,000	199 (24)	294 (24)	41 (22)
Between $10,000 and $20,000	147 (18)	210 (17)	33 (18)
Higher than $20,000	187 (23)	295 (24)	53 (28)

Data are mean ± SD or n (%). **Remaining percents are missing data,**

Table 2: Correlations between each axis and age, first 18 axes.

	Patients (first visit only)		Visits		Difference between round 1 and round 2	
Axis	Pearson R	p-value	Pearson R	p-value	Pearson R	p-value
1	0	0.79	-0.02	0.19	0.01	0.75
2	0.14	0	-0.12	0	0.10	0
3	0.06	0	-0.08	0	0.11	0
4	0.07	0	-0.05	0.01	-0.05	0.12
5	0.01	0.67	0.02	0.22	-0.09	0.01
6	-0.07	0	-0.07	0	-0.04	0.27
7	-0.09	0	0.08	0	-0.13	0
8	0	0.96	-0.01	0.57	-0.08	0.03
9	-0.08	0	-0.01	0.69	0.04	0.30
10	0	0.73	-0.10	0	-0.01	0.87
11	0.01	0.54	0.03	0.07	0.01	0.80
12	-0.01	0.96	-0.05	0	0.05	0.13
13	0	0	0.03	0.07	0.07	0.05
14	-0.08	0	0	0.81	0.02	0.52
15	0.06	0.51	0.05	0.01	-0.04	0.26
16	-0.01	0	-0.06	0	0.10	0.01
17	0.07	0	-0.03	0.11	0.06	0.07
18	-0.04	0.03	0	0.97	-0.05	0.16

Table 3: Characteristics of the first PCA axis

	Visits	Patients (first visit only)	Difference between round 1 and round 2
n	1230	818	187
Variance explained	8.7%	9.0%	9.9%
Top 20 loadings			
Hemoglobin	0.356 (6.38)	0.361 (6.40)	0.341 (5.76)
Hematocrit	0.311 (5.56)	0.321 (5.69)	0.242 (4.09)
Iron	0.304 (5.44)	0.294 (5.20)	0.247 (4.17)
Albumin	0.289 (5.17)	0.284 (5.03)	0.337 (5.69)
Saturated iron	0.247 (4.41)	0.237 (4.20)	0.219 (3.70)
Albumin/Globulin ratio	0.242 (4.34)	0.226 (4.01)	0.119 (2.01)
Protein-bound glucose	-0.212 (3.79)	-0.200 (3.55)	-0.116 (1.95)
MCH	0.209 (3.75)	0.185 (3.28)	0.129 (2.18)
Red blood cell dist. width	-0.195 (3.48)	-0.195 (3.46)	-0,075 (1.26)
Red blood cell count	0.190 (3.41)	0.222 (3.93)	0.278 (4.69)
Calcium	0.190 (3.40)	0.194 (3.44)	0.184 (3.11)
Creatinine	-0.183 (3.27)	-0.187 (3.32)	0.191 (3.23)
C-reactive protein	-0.181 (3.23)	-0.176 (3.11)	-0,085 (1.43)
Parathyroid hormone	-0.170 (3.04)	-0.185 (3.27)	0,034 (0.57)
Total bilirubin	0.134 (2.40)	0.147 (2.60)	0.159 (2.68)
Uric acid	-0.129 (2.30)	-0.111 (1.96)	0,076 (1.28)
25-hydroxy Vitamin D	0.124 (2.21)	0.132 (2.34)	0.109 (1.84)
Mean corpuscular hemoglobin conc.	0.118 (2.10)	0.107 (1.89)	0,066 (1.12)
Glucose	-0.115 (2.05)	-0.108 (1.91)	-0,086 (1.45)
Glycated hemoglobin	-0.113 (2.02)	-0.101 (1.78)	-0,063 (1.06)
Other biomarkers	28.25%	31.41%	38.62%

* Data are axis loadings with variance of the axis explained in parenthesis. Loadings are shown for the top 20 variables associated with the full data set (1230 visits).

Figure 1: Sum of the correlations between the i^{th} variable of the population principal component and the j^{th} variable of the random sample PCA – i.e., stability of each of the principal components under bootstraping procedures.

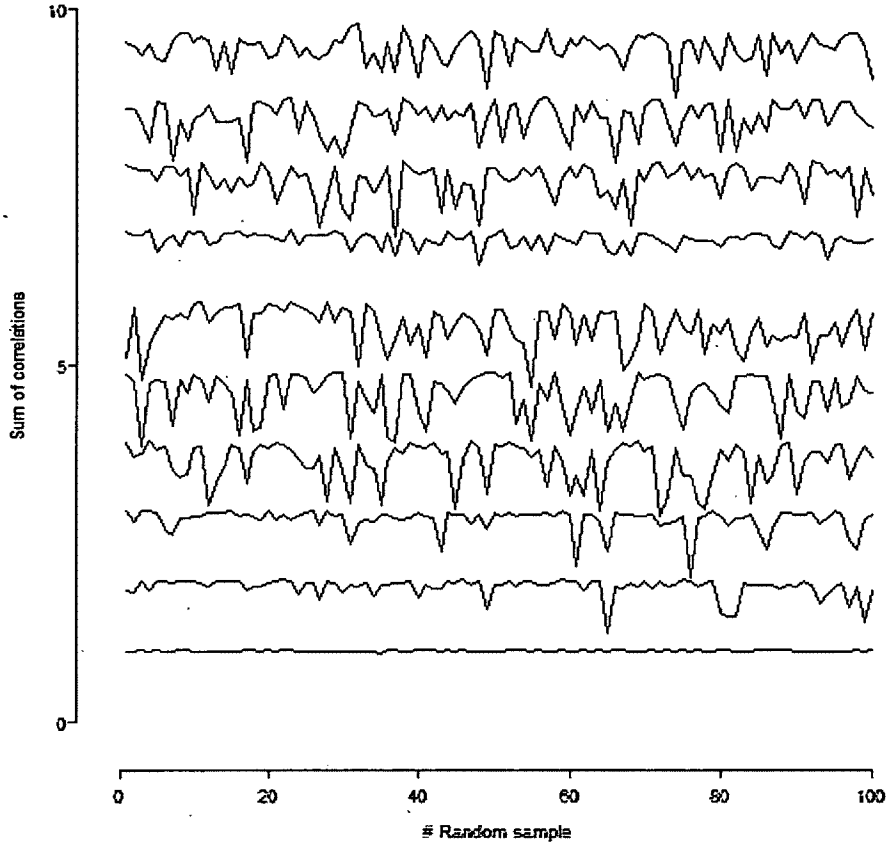


Figure 2: Sum of the correlations between the i^{th} variable of the population principal component and the j^{th} variable of the nonrandom sample PCA – i.e., stability of each of the principal components across the 13 subsamples.

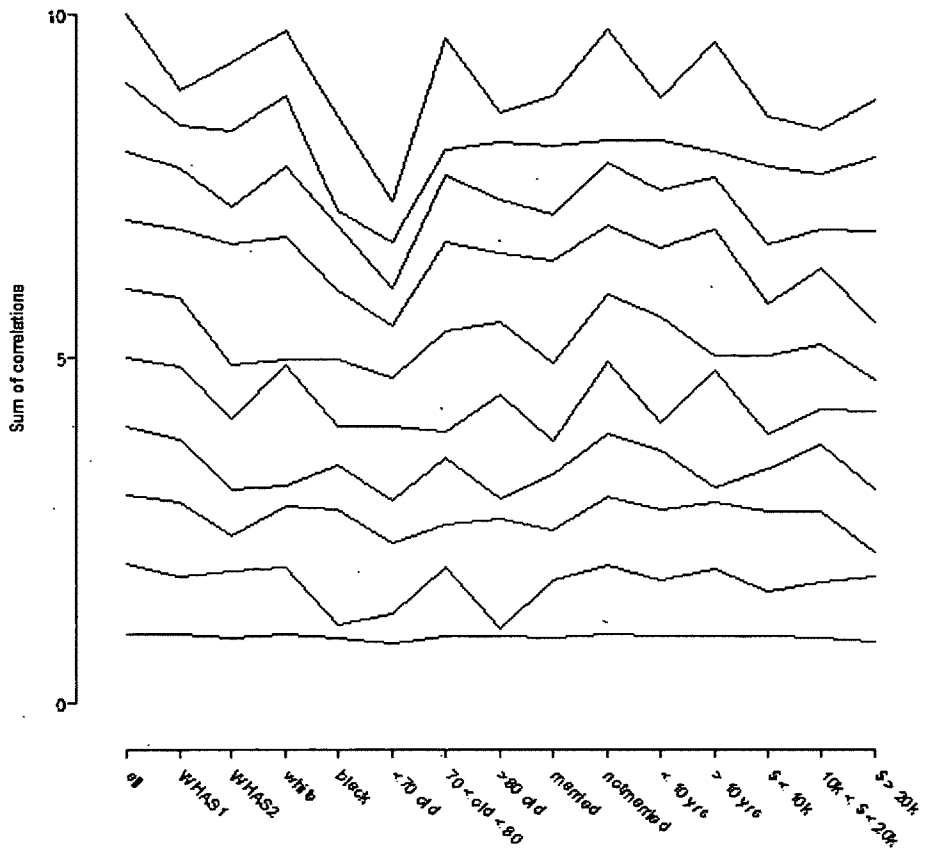
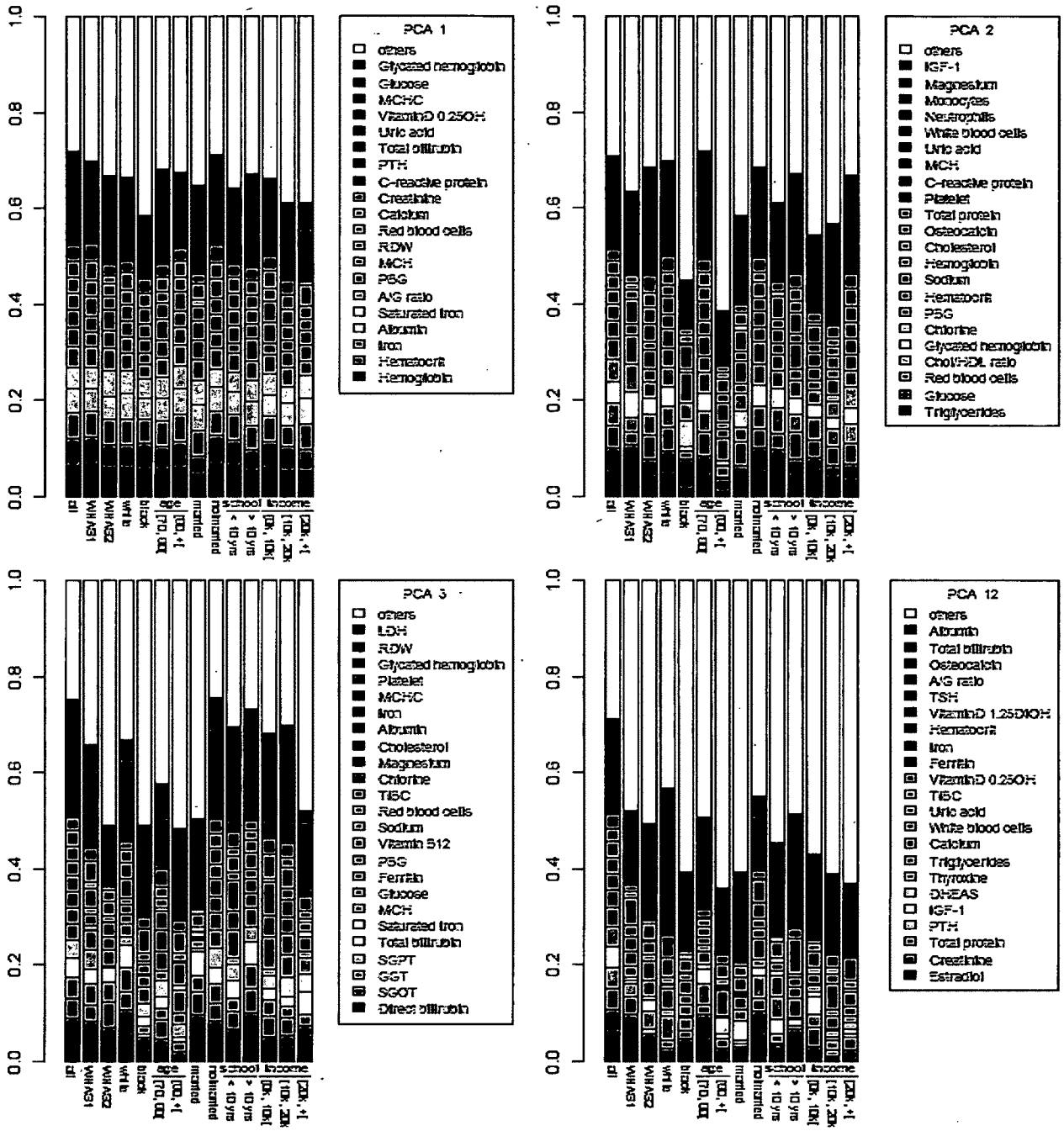


Figure 3: Composition of PCA axes 1, 2, 3, and 12 for each data subsample based on the 20 variables explaining the most variance in the axis as calculated from the full data set. Note that axis 1 is highly stable across subgroups, axes 2-3 are moderately stable, and axis 12 is highly unstable, with its composition changing markedly across subsamples.



** Abbreviations are : **A/G ratio** - Albumin/Globulin ratio, **DHEAS** - Dehydroepiandrosterone Sulfate, **GGT** - Gamma glutamyl transpeptidase, **IGF-1** - Insulin-like Growth Factor 1, **LDH** - Lactate dehydrogenase, **MCH** - Mean Corpuscular Hemoglobin, **MCHC** - Mean Corpuscular Hemoglobin Concentration, **PBG** - Protein-bound Glucose, **PTH** - Parathyroid hormone, **RDW** - Red blood cell Distribution Width, **SGOT** - Aspartate transaminase, **SGPT** - Alanine transaminase, **TIBC** - Total iron-binding capacity.*

Supplementary table: Characteristics of the biomarkers in all 1230 visits from 818 women aged 65 to 102 years old in Baltimore between 1993 and 2008.

	median	mean	sd	min	max	Correlation with age	
						R	p
Blood basics							
Hematocrit	39.2	39.03	3.582	25.5	50.6	-0.1554	<0.0001
Mean corpuscular hemoglobin	30.7	30.52	2.083	19.9	36.1	0.0584	0.0406
Mean corpuscular hemoglobin concentration	33.5	33.32	1.223	24.7	36.4	0.0224	0.4332
Red blood cells count	4.28	4.271	41.090	2.76	5.85	-0.1772	<0.0001
Red blood cells distribution width	13.8	14.08	1.368	11.6	23.9	0.0918	0.0013
Platelet	233	236.9	60.430	15	572	-0.0612	0.0318
Immune measures							
Basophil (%)	0.7	0.722	0.527	0	2.9	0.0018	0.9490
Eosophil (%)	2.6	3.135	2.189	0	16.4	0.0241	0.3989
Lymphocytes (%)	28.05	29.02	9.121	4.9	96	-0.0855	0.0027
Monocytes (%)	6.7	6.858	2.412	0	24.1	0.0796	0.0052
Neutrophil (%)	61.4	60.21	10.030	3	89.7	0.0578	0.0428
White blood cells count	6100	6416	3081	1700	95100	0.0643	0.0242
Hormones							
Dehydroepiandrosterone Sulfate	0.35	0.445	0.344	0.09	3	-0.1232	<0.0001
Estradiol	11.88	16.98	18.750	0.55	174.09	-0.1042	0.0003
Insulin-like Growth Factor 1	114.8	120.5	56.340	15	785.3	-0.1225	<0.0001
Parathyroid hormone	4	4.792	3.363	0.85	53.2	0.0518	0.0692
Thyroxine	7.8	7.88	1.806	2.9	16.8	-0.1318	<0.0001
Thyroid-stimulating hormone	1.6	2.11	2.441	0.02	39	0.0524	0.0660
Inflammatory measures							
C-reactive protein	3.1	6.191	8.910	2	111	-0.0157	0.5828
Interleukine 6	2.82	4.788	13.230	0.4	387.1	0.0717	0.0119
Ions							
Calcium	9.4	9.407	0.439	6.5	11.3	-0.1420	<0.0001
Chlorine	103	102.8	3.527	84	114	-0.0670	0.0188
Magnesium	2.01	1.994	0.199	1.1	2.63	0.0400	0.1607
Phosphorous	3.6	3.603	0.486	2.1	5.8	-0.0326	0.2533
Potassium	4.2	4.176	0.408	2.8	6	0.0199	0.4849
Sodium	140	139.7	2.602	126	148	-0.567	0.0467

Lipids							
Cholesterol/HDl ratio	4.3	4.474	1.492	2	12.7	-0.1040	0.0003
Cholesterol (%)	43	44.51	26.540	1	99	0.0124	0.6640
Triglycerides	136	156.16	92.440	26	807	-0.1195	<0.0001
Oxygen transport							
Ferritin	76.5	105.100	96.080	5.4	881	-0.0468	0.1006
Hemoglobin	13	13	1.194	8.1	16.5	-0.1492	<0.0001
Iron	79	81	27.340	17	274	-0.478	0.0936
Saturated iron	25	25.31	8.729	3.1	65.2	0.0289	0.3119
Total iron-binding capacity	324.5	326.700	49.530	193	548	-0.1474	<0.0001
Proteins, nitrogen metabolism, kidneys and liver							
Albumin/Globulin ratio	1.46	1.482	0.269	0.61	2.65	-0.0232	0.4154
Albumin	4.1	4.137	0.295	2.8	5.3	-0.2067	<0.0001
Alkaline phoshatase	82	87.32	35.160	30	666	0.0318	0.2650
Direct bilirubin	0.07	0.076	0.042	0.02	0.56		
Total bilirubin	0.42	0.465	0.228	0.11	3	0.1234	<0.0001
BUN/Creatinine ratio	17.3	17.77	4.828	6	37.8	0.1144	0.0001
Creatinine	1	1.037	0.322	0.6	6.1	0.1271	<0.0001
Gamma glutamyl transpeptidase	21	30.01	32.410	5	477	-0.0240	0.4010
Lactate dehydrogenase	173	176.4	34.310	67	450	0.0152	0.5955
Osteocalcin	5.65	9.73	10.950	0.5	157.5	0.0945	0.0009
Total protein	7	7	0.475	5.6	8.6	-0.1722	<0.0001
Aspartate transaminase	17.5	18.96	13.500	7	428	0.0423	0.1383
Alanine transaminase	14	15.84	14.460	2	387	-0.0796	0.0052
Uric acid	5.4	5.538	1.581	1.7	12.5	0.0596	0.0368
Sugar metabolism							
Glucose	99	115.2	53.940	46	549	-0.0529	0.0637
HbA1C	5.9	6.22	1.334	2.5	17.6	-0.0839	0.0032
Protein-bound Glucose	1.04	1.06	0.154	0.72	2.26	-0.0132	0.6448
Micronutrients and vitamins							
Folate	9.7	12.22	8.812	1.1	88.5	0.1242	<0.0001
Vitamin B12	438.5	497.4	299.090	104	5319	-0.0139	0.6269
25-hydroxy vitamin D	19	20.29	10.490	3	121.8	-0.0641	0.0245
1.25-dihydroxy vitamin D	39	40.22	14.030	5.2	116	-0.1452	<0.0001

Chapitre 5

CONCLUSION

5.1 Atteinte des objectifs

5.1.1 Le premier objectif

Le premier objectif de la recherche était de développer une routine d'analyse nous permettant d'analyser la structure d'une base de données en détail. Cet objectif a mené au développement d'une procédure regroupant plusieurs types d'analyse statistique, soient :

1. Analyse en composantes principales
2. Analyse parallèle de Horn
3. Algorithme de Daudin, Duby et Trecourt
4. Analyse des scores alternatifs

Grâce à cette routine, on est en mesure d'identifier quels sont les axes principaux gérant la base de données afin de les interpréter selon le contexte (analyse 1) et de les dénombrer (analyse 2). Nous sommes aussi capables d'évaluer la stabilité de la structure sur laquelle se basent les données et déterminer si la structure est un résultat généralisable à la population entière ou si au contraire ce n'est qu'une caractéristique de l'échantillon étudié (analyses 3 et 4).

5.1.2 Le second objectif

Le second objectif consistait à déterminer si le syndrome métabolique était un processus physiologique ou simplement une réunion de facteurs biologiques indépendants.

Suite à nos analyses, nous avons atteint notre objectif en indiquant qu'il existe bel et bien une structure gérant les cinq critères identifiés s'apparentant au syndrome métabolique.

En effet, la routine identifie clairement une première composante stable et presque identique au syndrome métabolique tel que défini dans la littérature médicale. En plus de correspondre à la structure du syndrome, le premier axe nous permet aussi de prédire efficacement les cas de maladies futures et confirme en ce sens l'utilité du syndrome métabolique.

Ainsi, de futurs travaux basés sur nos résultats pourraient amener la communauté scientifique ou médicale à construire un outil de prévention clinique permettant d'intervenir auprès des patients avant même qu'ils ne développent une maladie cardiovasculaire ou le diabète.

5.1.3 Le troisième objectif

Le dernier objectif nous amenait à utiliser la routine afin d'identifier si un processus semblable au vieillissement se cachait à travers le système de biomarqueurs humains.

Suite à nos analyses, l'atteinte du troisième objectif demeure toujours ambigüe puisque nous décelons bel et bien un processus régissant le système de biomarqueurs, mais celui-ci n'est pas relié au vieillissement.

En effet, la routine décèle trois composantes principales stables, mais celles-ci ne s'apparentent à aucun système connu de biomarqueurs ou de théories physiologiques connues. De plus, aucun processus ne semble être relié au vieillissement puisque seulement de faibles corrélations ont été détectées.

Ainsi, on doit considérer le troisième objectif comme partiellement achevé : nous n'avons pas trouvé de processus physiologique s'apparentant au vieillissement, mais avons pu distinguer un processus régissant la régulation des biomarqueurs étudiés.

5.2 Avenues de recherche futures

5.2.1 Concernant le syndrome métabolique

La base de données utilisée pour étudier le syndrome métabolique présente une grande homogénéité de la race des individus. En effet, plus de 95% des patients étudiés sont Caucasiens, il est donc impératif de répéter les analyses sur une population plus diversifiée ou composée majoritairement d'une autre race (afro-américains, latino-américains, asiatiques, etc.) pour affiner les résultats.

Le même problème se pose concernant la santé relative des individus puisque l'échantillon étudié provient des données administratives d'un hôpital. Il est donc naturel de croire que la plupart des patients étudiés étaient malades au moment de la collecte des données. On devrait donc répéter les analyses sur une cohorte plus en santé pour affiner les résultats.

5.2.2 Concernant le vieillissement

Les résultats concernant le processus physiologique du vieillissement, quant à eux, sont un peu plus flous. Il y a donc plusieurs avenues de recherche qui en découlent.

Premièrement, l'échantillon étudié est encore une fois assez homogène, notamment au niveau du sexe et de l'âge. En effet, toutes les patientes étudiées sont des femmes âgées de 65 ans et plus, donc aucune interprétation possible pour les hommes ou les jeunes femmes.

Le manque de variabilité au niveau de l'âge pourrait aussi être à l'origine de la faible corrélation unissant l'âge au processus physiologique identifié dans l'étude. On devrait donc mettre en priorité l'étude d'une cohorte plus hétérogène afin de confirmer les résultats tirés de l'étude. Ces analyses supplémentaires sont d'autant plus importantes si on considère aussi le fait que les femmes ayant participé à l'étude étaient ménopausées.

En ce qui concerne la corrélation, il serait aussi pertinent de reprendre les analyses et de tenter d'identifier une corrélation non linéaire entre le processus identifié et l'âge des individus.

Il serait aussi pertinent d'étudier le processus physiologique du vieillissement à l'aide de méthodes différentes comme des modèles d'équations structurelles ou des analyses de distances multivariées (distance de Mahalanobis par exemple) entre les biomarqueurs et l'âge des patients pour confirmer ou infirmer les résultats.

Enfin, il serait intéressant de déterminer si le processus physiologique identifié par les analyses est en mesure de prédire les risques de maladies un peu comme en est capable le syndrome métabolique en ce qui concerne les risques de maladies cardiovasculaires ou de diabète.

Bibliographie

[1] M.S. Bartlett. Tests of significance in factor analysis. *British Journal of Psychology*, 3 :77–85, 1950.

[2] E. Beltrami. Sulle funzioni bilineari. *Giornale di Mathematiche di Battaglini*, 11 :691–702, 1873.

[3] R.B. Catell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1 :245–276, 1966.

[4] R.S. PAdwal et A.M. Sharma. Prevention of cardiovascula disease : obesity, diabetes and the metabolic syndrome. *Canadian Journal of Cardiology*, 26 :18–20, 2010.

[5] J.J. Daudin C. Duby et P.Trecourt. Stability of Principal Component Analysis Studied by the Bootstrap Method. *Statistics*, 19 :241–258, 1988.

[6] J.C. Hayton D.G. Allen et V.Scarpello. Factor Retention Decisions in Explanatory Factor Analysis : A Tutorial on Parallel Analysis. *Organizational Research Methods*, 7 :191–205, 2004.

[7] R.A. Fisher et W.A. Mackenzie. Studies in crop variation ii. the manurial response of different potato varieties. *Journal of Agricultural Science*, 13 :311–320, 1923.

[8] W.R. Zwick et W.F. Velicer. Comparison of Five Rules for Determining the Number of Components to Retain. *Psychological Bulletin*, 99 :432–442, 1986.

[9] J.L. Horn. A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika*, 30 :179–185, 1965.

[10] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24 :417–441, 1933.

[11] I.T. Joliffe. *Principal component analysis*. Springer, 2nd edition, 2002.

[12] M.C. Jordan. Mémoire sur les Formes Bilinéaires. *Journal de mathématiques Pures et Appliquées*, 19 :35–54, 1874.

[13] H.F. Kaiser. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20 :141–151, 1960.

[14] K. Pearson. O Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2 :559–572, 1901.