Matematisk statistik
Stockholms universitet

Kanonisk korrelationsanalys av gendata och kliniska data på patienter med åderförkalkning i halspulsådern

Niko Asyabani

Examensarbete 2005:1

Postadress:

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm
Sverige

Internet:

http://www.math.su.se/matstat

Matematisk statistik
Stockholms universitet
Examensarbete **2005:1**,
http://www.math.su.se/matstat

Kanonisk korrelationsanalys av gendata och kliniska data på patienter med åderförkalkning i halspulsådern

Niko Asyabani*

januari 2005

Sammanfattning

Det här arbetet har gjorts på uppdrag av Karotisprojektet som har som huvuduppgift att undersöka sambandet mellan inflammation och åderförkalkning i halspulsådern. Syftet med detta arbete är att undersöka om det föreligger samband mellan nio kliniska labparametrar (CRP, S-Kolesterol, LPK, Fibrinogen, Tg, HDL, LDL, Hb, EVF) och sju gener (CD1a, CD1b, CD1c, CD1d, CD1e, Thromboxan, Vα24) hos patienter med åderförkalkning i halspulsådern. Mätningarna har gjorts på 54 patienter varav 37 är män och 17 är kvinnor. CD1-generna och Vα24 är gener som uttrycks i de vita blodkropparna som ingår i immunförsvaret. Anledningen till att man vill undersöka dessa gener är att inflammation framkallas i kroppen när immunförsvaret överaktiveras. Thromboxan är gener som uttrycks i blodplättarna som ingår i blodets koaguleringsprocess. Den metod som har använts för att undersöka sambandet är kanonisk korrelationsanalys. Resultaten av denna metod visar att det inte finns något signifikant samband mellan dessa gener och de kliniska labparametrarna. Däremot finns det indikationer på att variablerna Thromboxan, Fibrinogen och EVF har störst betydelse. Det krävs dock en utökning av patientdata för att kunna belägga detta. Inga signifikanta samband kunde heller påvisas varken för män eller kvinnor när de undersöktes som två separata grupper, möjligen på grund av små stickprov.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige. E-post: md01nas@math.su.se Handledare: Mikael Andersson.

Abstract

The origin of this paper is the Carotis project whose main purpose is to investigate the connection between inflammation and arteriosclerosis in the carotid artery. The objective of this work is to examine if there is a connection between nine clinical labparameters (CRP, S-Cholesterol, LPK, Fibrinogen, Tg, HDL, LDL, Hb, EVF) and seven genes (CD1a, CD1b, CD1c, CD1d, CD1e, Thromboxane, Vα24) for patients with arteriosclerosis in the carotid artery. The CD1-genes and Vα24 are expressed in the white blood corpuscles that are a part of the immune system. The choice of these genes is coroborated by the fact that inflammation is induced in the body when the immune system is overactivated. Thromboxane is expressed in the blood-plates that are a part of the coagulation process of the blood. The method being used to investigate the connection is Canonical correlation analysis. The results of this method show that there is no significant connection between the genes and the clinical labparameters. However, there are indications that the variables Thromboxane, Fibrinogen and EVF are of greatest importance. In any ambition to stress this, the patient data have to be increased. Finally, there are still no significant connections between the genes and the clinical labparameters when men and women are investigated as two separate groups, perhaps because of small samples.

Förord

Jag skulle vilja tacka min handledare Mikael Andersson på Matematiska institutionen på Stockholms Universitet för att ha varit till stor hjälp under arbetets gång.

Jag skulle också vilja tacka min handledare Gabrielle Berne och Anders Gabrielsen på CMM på Karolinska Institutet som har hjälpt mig med den medicinska delen.

Innehållsförteckning

1 Introduktion

I Sverige beror hälften av dödsfallen på hjärt- och kärlsjukdomar och dessa sjukdomar förväntas vara den vanligaste dödsorsaken omkring år 2010 till 2020. I de flesta fall beror sjukdomen på åderförkalkning. Genom att minska de största riskfaktorerna (diabetes, rökning, högt blodtryck och högt kolesterolvärde) har dödlighet förenad med åderförkalkning minskat. Men dessa riskfaktorer är inte de enda orsakerna till åderförkalkning. Kliniska studier har nyligen visat att det finns ett samband mellan inflammationer i kroppen och åderförkalkning, speciellt hos diabetiker.

Karotisprojektet[1], som är ett samarbete mellan Kardiovaskulära[2] forskningslaboratoriet vid Centrum för Molekylär Medicin (CMM) och Kärlkirurgiska kliniken, båda vid Karolinska sjukhuset och Karolinska institutet, har som huvuduppgift att undersöka sambandet mellan inflammation och åderförkalkning i halspulsådern.

Syftet med detta arbete, som har gjorts på uppdrag av Karotisprojektet, är att undersöka om det föreligger samband mellan nio kliniska labparametrar (CRP, S-Kolesterol, LPK, Fibrinogen, Tg, HDL, LDL, Hb, EVF) och sju gener (CD1a, CD1b, CD1c, CD1d, CD1e, Thromboxan, Vα24) hos karotispatienter (Patienter med åderförkalkning i halspulsådern). Detta för att se hur labparametrar och gener samvarierar det vill säga vad som händer med genernas proteinuppsättning när t ex kolesterolvärdet är högt. Kan man hitta ett samband mellan gener och labparametrar? Alla sexton variabler antar kontinuerliga värden.

Datasetet har tagits fram av CMM. Mätningarna av labparametrar och gener har gjorts på 54 patienter som har opererats för åderförkalkning i halspulsådern. Alla mätningar har gjorts i samband med operationen.

1 Karotis är den medicinska benämningen på halspulsådern.

2 Medicinsk term som har med hjärtat och blodkärlen att göra.

2 Medicinsk bakgrund

2.1 Labparametrar

CRP (C-reaktivt protein)
CRP är ett protein som bildas i ökad omfattning vid akuta infektioner och inflammatoriska sjukdomar. Halten CRP kan bestämmas av blodprov.

Blodlipider (blodfett) är fettliknande ämnen i blodet och kan kemiskt delas in i tre grupper: triglycerider (neutralfett), kolesterol och fosfolipider.
Lipoproteiner är partiklar med uppgift att transportera fettämnen i blodet.

S-Kolesterol
I kroppen förekommer kolesterol i alla vävnader, främst i hjärna och ryggmärg. Vid förhöjd kolesterolhalt i blodet ökar risken för åderförkalkning.
S-kolesterol är kolesterolkoncentrationen i serum. Serum fås fram genom att levrat blod som håller på att dra sig samman pressas. Vid pressning försvinner cellerna och en del proteiner.

Tg (Triglycerider)
Kallas även för neutralfett och är en grupp av blodfetter.

LDL (Low density lipoproteins)
LDL är de mest kolesterolrika lipoproteinerna. Omkring 80 % av blodets kolesterol transporteras normalt som LDL-partiklar.

HDL (High density lipoproteins)
HDL är relativt rik på proteiner och fosfolipid och anses ha förmåga att ta upp överskott av kolesterol från kroppens vävnader. Denna mekanism kan bl.a. ha betydelse för att förhindra åderförkalkning.

LPK (Leukocyter partikel koncentration)
Leukocyter är de vita blodkropparna.

Fibrinogen

Lösligt protein i blodplasma. Fibrinogen tillhör gruppen akutfasproteiner, dvs halten av fibrinogen stiger snabbt vid inflammation.

Hb

Hemoglobin, *Hb*, är ett syrebindande protein som finns i hög halt i de röda blodkropparna (erytrocyterna) och ger blodet dess röda färg. Hemoglobinet har förmåga att ta upp och binda syre och att avge det. Koncentrationen av hemoglobin i blodet uppgår normalt till 132-163 g/l hos män och 116-148 g/l hos kvinnor.

EVF (Erytrocyternas volymfraktion)

De röda blodkropparnas (erytrocyternas) procentuella andel av blodet. Detta värde varierar normalt hos män mellan 39 % och 49 % (EVF 0,39-0,49) samt hos kvinnor mellan 37 % och 44 % (EVF 0,37-0,44).

Informationen är hämtad från www.ne.se

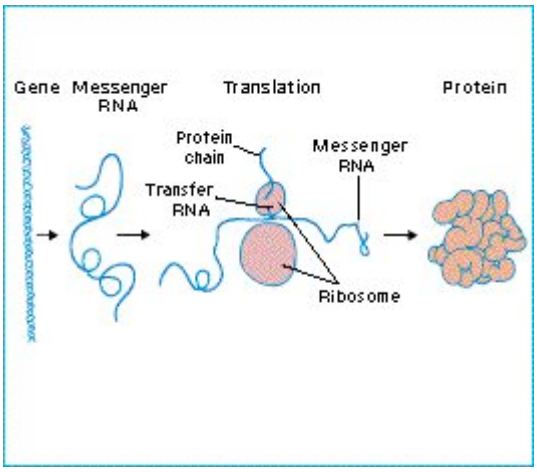
2.2 Gener

Den minsta levande enheten i människokroppen är cellen. De arbetsuppgifter cellen utför beror på arvsanlagen som finns i cellkärnan. Arvsanlagen, eller gener som de också kallas, innehåller information om hur vi ser ut och fungerar och går i arv från generation till generation.

En bland många molekyler i cellen är DNA-molekylen som finns i cellkärnan. DNA-molekylen består av fyra olika kvävebaser. De kallas för A, T, C och G efter första bokstaven i deras namn. Kvävebaserna uppträder ett mycket stort antal gånger i en DNA-molekyl. Skulle man titta på en gen är det antalet, ordningen och kombinationen av kvävebaser som avgör vad det är för gen. En gen är alltså en del av en DNA-molekyl och innehåller tusentals kvävebaser. En DNA-molekyl i sin tur kan innehålla tusentals gener.

Proteiner är andra molekyler som finns i cellen. En proteinmolekyl är en kedja av hopkopplade aminosyror. Det finns 20 olika aminosyror som kan kopplas ihop i olika ordningsföljder. Ordningsföljden avgör hur proteinet ser ut dvs vilken typ den är. Protein används främst som byggmaterial i cellerna och utför nästan alla processer i en levande varelse och är avgörande för hur vi ser ut och fungerar.

När en cell bildar protein är det generna (kvävebasernas ordningsföljd i ett avsnitt av DNA-molekylerna) som bestämmer i vilken ordning aminosyrorna ska kopplas ihop. Med andra ord är det DNA-molekyler som bestämmer vilka proteiner som bildas. Detta görs genom att delar av cellens DNA-molekyler kopieras till en mRNA-molekyl (MessengerRNA). mRNA-molekylen innehåller då information om sammansättningen av det protein som ska bildas. Hur proteinet ser ut beror på vilka DNA-molekyler som har kopierats.



Figur 1: Illustration över proteinsyntes (Hämtad från http://www.blackwellpublishing.com/ridley/images/gene_information_transfer.jpg)

Genvärdena i denna undersökning har tagits från mRNA-molekylen (som även kallas för genuttrycket).

CD1-generna och Vα24 är gener som uttrycks i de vita blodkropparna som ingår i immunförsvaret. Anledningen till att man vill undersöka dessa gener är att inflammation framkallas i kroppen när immunförsvaret överaktiveras.

Thromboxan är gener som särskilt uttrycks i blodplättarna (en av blodets beståndsdelar) som ingår i blodets koagulationsprocess. Koagulationens viktigaste uppgift är att stoppa blödningar som uppkommer vid skador på blodkärlen.

Genvärden har bestämts med Taqman-metoden. Det värde som fås normaliseras först mot en ”housekeeping” gen (Cyclophilin A) och sedan mot en kontrollvävnad från en frisk person.

3 Multivariat analys

Till multivariat analys räknas de statistiska metoder som analyserar stickprov från flerdimensionella fördelningar, särskilt multivariat normalfördelning. Täthetsfunktionen för multivariat normalfördelning är en generalisering av täthetsfördelningen för normalfördelning. Täthetsfördelningen för normalfördelningen med väntevärdet μ och variansen σ^2 är:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x-\mu)/\sigma]^2/2} \qquad -\infty < x < -\infty$$

Täthetsfunktionen för multivariat normalfördelning är:

$$f_{\mathbf{X}}(\mathbf{x}) = \left(\frac{1}{2\boldsymbol{\pi}}\right)^{\frac{n}{2}} \frac{1}{\sqrt{\det \Sigma}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \qquad -\infty < x_i < \infty,\ i = 1,2,\cdots,n$$

där $\mathbf{X}$ är en n-dimensionell vektor med väntevärdesvektorn $\boldsymbol{\mu} = E(\mathbf{X})$ och kovariansmatrisen $\Sigma = E(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'$.

Det finns flera olika metoder inom multivariat analys. För att veta vilken metod man ska använda sig av brukar man titta på några saker:
Kan man definiera variablerna som respons- och förklarande variabler? Om denna indelning kan göras, hur många respons- respektive förklarande variabler finns det? Är variablerna diskreta eller kontinuerliga?

I vårt fall har vi sju gener och nio labparametrar och vi ska se hur variablerna i de två mängderna samvarierar. Det som är intressant är variationen mellan de två mängderna och inte variationen inom respektive mängd. Den metod som passar bäst till vårt dataset är kanonisk korrelationsanalys. Vi ska inte försöka förklara den ena mängden med hjälp av den andra mängden, utan bara undersöka hur de två mängderna samvarierar. Man brukar ändå lite slarvigt kalla ena mängden för responsvariabler och den andra för förklarande variabler. Denna benämning kommer att användas fritt i fortsättningen av detta arbete. Det är viktigt att notera att resultatet blir samma oavsett om man väljer generna som responsvariabler eller labparametrarna.

4 Kanonisk korrelationsanalys

Kanonisk korrelationsanalys undersöker relationen mellan två mängder av variabler. Den ena mängden innehåller ”responsvariabler” och den andra mängden innehåller ”förklarande variabler”. Först bestäms en kanonisk funktion. En kanonisk funktion består av två kanoniska variabler. Kanoniska variabler är konstruerade så att korrelationen mellan dem är maximal. Ena kanoniska variabeln är en linjärkombination av responsvariablerna och den andra en linjärkombination av de förklarande variablerna. Detta görs med hjälp av kanoniska vikter (kanoniska koefficienter). De transformerar originalvariablerna så att korrelationen mellan responsmängden och förklarandemängden i en kanonisk funktion blir maximal. Storleken på vikterna anger hur betydelsefulla originalvariablerna är i den kanoniska funktionen.

Korrelationen mellan de två kanoniska variablerna i en funktion kallas för kanonisk korrelation. Den första funktionen har den högsta kanoniska korrelationen. Sedan bestäms nästa funktion som den som har den högsta korrelationen bland resterande funktioner som är okorrelerade med första funktionen. Tredje kanoniska funktionen bestäms som den som har högsta korrelationen bland resterande funktioner som är okorrelerade med de två första och så vidare.

Kvadrerade kanoniska korrelationer kallas för kanoniska rötter eller egenvärden. Kanoniska rötter ger en skattning av andelen delad varians mellan två kanoniska variabler i en kanonisk funktion.

Antalet kanoniska funktioner motsvarar minimum av antal responsvariabler och antal förklarande variabler.

4.1 Matematisk modell

Allmän definition:

Antag att $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_r \end{bmatrix}$ d v s $\mathbf{X}$ är en r-dimensionell stokastisk vektor.

Vi definierar medelvärdet $\mu_i = E(X_i)$ där $i = 1,2,\cdots,r$.

Då är medelvärdesvektorn $E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_r) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_r \end{bmatrix} = \boldsymbol{\mu}$

Och kovariansmatrisen

$$\sum = E(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})' = E\left(\begin{bmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ \vdots \\ X_r-\mu_r \end{bmatrix}\begin{bmatrix} X_1-\mu_1, X_2-\mu_2,\ldots, X_r-\mu_r \end{bmatrix}\right) =$$

$$= \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \cdots & Cov(X_1,X_r) \\ Cov(X_2,X_1) & Var(X_2) & \cdots & Cov(X_2,X_r) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_r,X_1) & Cov(X_r,X_2) & \cdots & Var(X_r) \end{bmatrix}$$

Antag att $\mathbf{c}$ är en r-dimensionell vektor med konstanter d v s $\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_r \end{bmatrix}$.

Då har den linjära kombinationen $\mathbf{c}'\mathbf{X} = c_1X_1 + \cdots + c_rX_r$ följande egenskaper:

Medelvärdesvektorn = $E(\mathbf{c}'\mathbf{X}) = \mathbf{c}'\boldsymbol{\mu}$

Variansen = $Var(\mathbf{c}'\mathbf{X}) = \mathbf{c}'Cov(\mathbf{X},\mathbf{X})\mathbf{c} = \mathbf{c}'\sum\mathbf{c}$

Vi ska undersöka sambandet mellan två mängder. Första mängden innehåller p stycken variabler och representeras av den p-dimensionella stokastiska vektorn $\mathbf{X}^{(1)} = \begin{bmatrix} X_1^{(1)} \\ \vdots \\ X_p^{(1)} \end{bmatrix}$. Den andra mängden innehåller q stycken variabler och betecknas med den q-dimensionella stokastiska vektorn $\mathbf{X}^{(2)} = \begin{bmatrix} X_1^{(2)} \\ \vdots \\ X_q^{(2)} \end{bmatrix}$. Vi antar att $p \le q$.

Vi inför följande beteckningar för de stokastiska vektorerna $\mathbf{X}^{(1)}$ och $\mathbf{X}^{(2)}$:

$E(\mathbf{X}^{(1)}) = \boldsymbol{\mu}^{(1)}$

$E(\mathbf{X}^{(2)}) = \boldsymbol{\mu}^{(2)}$

$Var(\mathbf{X}^{(1)}) = Cov(\mathbf{X}^{(1)},\mathbf{X}^{(1)}) = \sum_{11}$

$Var(\mathbf{X}^{(2)}) = Cov(\mathbf{X}^{(2)},\mathbf{X}^{(2)}) = \sum_{22}$

$Cov(\mathbf{X}^{(1)},\mathbf{X}^{(2)}) = \sum_{12} = \sum_{21}'$

$Cov(\mathbf{X}^{(2)},\mathbf{X}^{(1)}) = \sum_{21} = \sum_{12}'$

Sammanslagning av $\mathbf{X}^{(1)}$ och $\mathbf{X}^{(2)}$ benäms som $\mathbf{X}$ och ser ut som följer:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix} = \begin{bmatrix} X_1^{(1)} \\ X_2^{(1)} \\ \vdots \\ X_p^{(1)} \\ X_1^{(2)} \\ X_2^{(2)} \\ \vdots \\ X_q^{(2)} \end{bmatrix}$$

dess medelvärdesvektor:

$$\boldsymbol{\mu} = E(\mathbf{X}) = \begin{bmatrix} E(\mathbf{X}^{(1)}) \\ E(\mathbf{X}^{(2)}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}$$

och dess kovariansmatris:

$$\sum = E(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})' = \begin{bmatrix} E(\mathbf{X}^{(1)}-\boldsymbol{\mu}^{(1)})(\mathbf{X}^{(1)}-\boldsymbol{\mu}^{(1)})' & E(\mathbf{X}^{(1)}-\boldsymbol{\mu}^{(1)})(\mathbf{X}^{(2)}-\boldsymbol{\mu}^{(2)})' \\ E(\mathbf{X}^{(2)}-\boldsymbol{\mu}^{(2)})(\mathbf{X}^{(1)}-\boldsymbol{\mu}^{(1)})' & E(\mathbf{X}^{(2)}-\boldsymbol{\mu}^{(2)})(\mathbf{X}^{(2)}-\boldsymbol{\mu}^{(2)})' \end{bmatrix} = \begin{bmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{bmatrix}$$

Kovariansen mellan par av variabler från olika mängder (kovariansen mellan en gen och en labparameter), en variabel från $\mathbf{X}^{(1)}$ och en variabel från $\mathbf{X}^{(2)}$, finns i $\sum_{12}$ som är identisk med $\sum_{21}'$. Syftet med kanonisk korrelationsanalys är att beräkna sambandet mellan $\mathbf{X}^{(1)}$ och $\mathbf{X}^{(2)}$. Av intresse är att beräkna korrelationen mellan linjära kombinationen av variablerna i de två mängderna.

Sätt

$$U = \mathbf{a}'\mathbf{X}^{(1)}$$
$$V = \mathbf{b}'\mathbf{X}^{(2)}$$

där $\mathbf{a}$ och $\mathbf{b}$ är koefficientvektorerna $\mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix}$ och $\mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_q \end{bmatrix}$.

$$Var(U) = \mathbf{a}'Cov(\mathbf{X}^{(1)}, \mathbf{X}^{(1)})\mathbf{a} = \mathbf{a}'\sum_{11}\mathbf{a}$$
$$Var(V) = \mathbf{b}'Cov(\mathbf{X}^{(2)}, \mathbf{X}^{(2)})\mathbf{b} = \mathbf{b}'\sum_{22}\mathbf{b}$$
$$Cov(U,V) = \mathbf{a}'Cov(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})\mathbf{b} = \mathbf{a}'\sum_{12}\mathbf{b}$$
$$Cov(V,U) = \mathbf{b}'Cov(\mathbf{X}^{(2)}, \mathbf{X}^{(1)})\mathbf{a} = \mathbf{b}'\sum_{21}\mathbf{a}$$

Vi söker koefficientvektorer $\mathbf{a}$ och $\mathbf{b}$ så att

$$Corr(U,V) = \frac{\mathbf{a}'\sum_{12}\mathbf{b}}{\sqrt{\mathbf{a}'\sum_{11}\mathbf{a}}\sqrt{\mathbf{b}'\sum_{22}\mathbf{b}}}$$

blir så stor som möjligt. Koefficientvektorerna **a** och **b** är våra kanoniska vikter.

Första kanoniska funktionen är det par av linjära kombinationer U_1, V_1 som har den högsta korrelationen.

Den andra kanoniska funktionen är det par av linjära kombinationer U_2, V_2 som har den högsta korrelationen bland alla de par som är okorrelerade med det första paret av kanoniska variabler.

Den k:te kanoniska funktionen är det par av linjära kombinationer U_k, V_k som har den högsta korrelationen bland de par som är okorrelerade med de tidigare k-1 paren av kanoniska variabler. Korrelationen mellan det k:te paret av linjära kombinationen U_k, V_k (eller k:te paret av kanoniska variabler) kallas för den k:te kanoniska korrelationen.

Påståenden:

$$\max_{a,b} Corr(U,V) = \rho^*_1$$

som fås genom de linjära kombinationerna (första paret av kanoniska variabler)

$$U_1 = \underbrace{\mathbf{e}_1'\sum_{11}^{-1/2}}_{\mathbf{a}_1'}\mathbf{X}^{(1)} \quad \text{och} \quad V_1 = \underbrace{\mathbf{f}_1'\sum_{22}^{-1/2}}_{\mathbf{b}_1'}\mathbf{X}^{(2)}$$

Det k:te paret av kanoniska variabler där k = 2, 3, ..., p är

$$U_k = \underbrace{\mathbf{e}_k'\sum_{11}^{-1/2}}_{\mathbf{a}_k'}\mathbf{X}^{(1)} \quad \text{och} \quad V_k = \underbrace{\mathbf{f}_k'\sum_{22}^{-1/2}}_{\mathbf{b}_k'}\mathbf{X}^{(2)}$$

och som maximerar $Corr(U_k,V_k) = \rho^*_k$ bland de kanoniska funktioner som är okorrelerade med de tidigare 1, 2, ..., k-1 kanoniska funktioner.

Här är $(\rho^*_1)^2 \geq (\rho^*_2)^2 \geq \cdots \geq (\rho^*_p)^2$ egenvärden av matrisen $\sum_{11}^{-1/2} \sum_{12} \sum_{22}^{-1} \sum_{21} \sum_{11}^{-1/2}$ och $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_p$ är de tillhörande p-dimensionella normerade egenvektorerna.

$(\rho^*_1)^2, (\rho^*_2)^2, \cdots, (\rho^*_p)^2$ är också de p största egenvärdena av matrisen $\sum_{22}^{-1/2} \sum_{21} \sum_{11}^{-1} \sum_{12} \sum_{22}^{-1/2}$ med tillhörande q-dimensionella normerade egenvektorerna $\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_p$. Varje $\mathbf{f}_i$ är proportionell mot $\sum_{22}^{-1/2} \sum_{21} \sum_{11}^{-1/2} \mathbf{e}_i$. Notera att vi har antagit att $p \leq q$.

De kanoniska variablerna har följande egenskaper:

$$Var(U_k) = Var(V_k) = 1$$
$$Cov(U_k, U_l) = Corr(U_k, U_l) = 0 \qquad k \neq l$$
$$Cov(V_k, V_l) = Corr(V_k, V_l) = 0 \qquad k \neq l$$
$$Cov(U_k, V_l) = Corr(U_k, V_l) = 0 \qquad k \neq l$$

för $k, l = 1, 2, \cdots, p$.

Bevis för påståendet och utförligare beräkningar finns i Applied Multivariate Statistical Analysis av Richard A. Johnson och Dean W. Wichern. I beviset antar man att matriserna $\sum_{11}$ och $\sum_{22}$ är ickesingulära. Med det menas att matriserna ska ha full rang dvs alla rader (eller kolonner) ska vara linjärt oberoende.

4.2 Antaganden för kanonisk korrelation

4.2.1 Normalitet

En viktig antagande för testerna i kanonisk korrelation (T ex F-test) är normalitet. Med det menas att variablerna ska vara multivariat normalfördelade. Eftersom det är svårt att testa det brukar man nöja sig med att undersöka om marginalfördelningen för varje variabel är lika med normalfördelningen. Ett sätt att undersöka normaliteten hos variablerna är grafiskt t ex

genom att rita histogram eller normalfördelningsplot. Ett annat sätt är att använda sig av de test som finns t ex Shapiro-Wilks normalitetstest. Små värden på teststatistikan

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

visar att stickprovet inte är normalfördelat, där $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ är det ordnade stickprovet och $a_1, a_2, \ldots, a_n$ är konstanter som har skapats av medelvärdet, variansen och kovariansen av en ordningsstatistika av ett normalfördelat stickprov av storlek n enligt:

$$\mathbf{a}' = \mathbf{m}'\mathbf{v}^{-1}\left[\left(\mathbf{m}\mathbf{v}^{-1}\right)\left(\mathbf{v}^{-1}\mathbf{m}\right)\right]^{-1/2}$$

$$\mathbf{m} = E\begin{bmatrix} X_{(1)} \\ X_{(2)} \\ \vdots \\ X_{(n)} \end{bmatrix} \qquad \mathbf{v} = Var\begin{bmatrix} X_{(1)} \\ X_{(2)} \\ \vdots \\ X_{(n)} \end{bmatrix}$$

4.2.2 Linjaritet

Linjaritet är när sambandet mellan två variabler är linjärt. Det vanligaste sättet att undersöka om linjaritet föreligger är att göra scatterplottar. I scatterplottar plottas variablerna parvis mot varandra. Mönstret av punkterna representerar sambandet mellan variablerna. Om punkterna är samlade längs en rät linje tyder det på linjärt samband eller korrelation. Om punkterna bildar en punktsvärm utan något mönster är det tecken på att korrelation saknas. Kanonisk korrelation är det linjära sambandet mellan två kanoniska variabler. Om något av sambanden inte är linjärt upptäcks inte sambandet.

4.2.3 Ej Kolinjaritet/Multikolinjaritet

Kolinjaritet råder när korrelationskoefficienten är 1 mellan två förklarande variabler och ingen kolinjaritet när korrelationskoefficienten är 0 (I vårt gäller detta även för responsvariabler). Om starka samband förekommer mellan två förklarande variabler kallar man det för nästan kolinjaritet. Om en förklarande variabel är starkt korrelerad med fler än en förklarande variabel kallas det för nästan multikolinjaritet.

För att testa kolinjaritet brukar man titta på flera saker. Korrelationskoefficienten mäter det linjära sambandet mellan två variabler. Men för att undersöka om fler än två variabler är korrelerade brukar man titta på toleransen och variansinflationen. Tolerans definieras som $1-R^2$ och beräknas för alla förklarande variabler där R^2 är förklaringsgraden för den förklarande variabeln predikterat av resten av förklarande variablerna. Förklaringsgraden är den del av variansen hos responsvariabeln som förklaras av regressorena. Om toleransen är nära noll indikerar det att variabeln beskrivs väldigt bra av de förklarande variablerna. (När man undersöker multikolinjaritet hos responsvariablerna beräknar man på samma sätt R^2 för alla responsvariabler vilket är förklaringsgraden för responsvariabeln predikterat av resten av responsvariabler).

Variansinflation är $\frac{1}{1-R^2}$ dvs inversen av toleransen. Höga värden på variansinflation visar att kolinjaritet eller multikolinjariet råder bland variablerna. En tumregel är att variansinflation över 10 visar på kolinjaritet eller multikolinjaritet.

Kolinjaritet/multikolinjaritet (eller nästan kolinjaritet/multikolinjaritet) måste åtgärdas eftersom den gör anpassningen mindre pålitlig. I beviset för kanonisk korrelationsanalys antar man att matriserna $\sum_{11}$ och $\sum_{22}$ är ickesingulära dvs alla kolonner är linjärt oberoende. Kolinjaritet/multikolinjaritet (eller nästan kolinjaritet/multikolinjaritet) visar på att två eller fler kolonner är linjärt beroende (eller nästan linjärt beroende).

4.3 Datatransformationer

Datatransformation är vanligt för att uppfylla sina antaganden (som har beskrivit i avsnitt 4.2). De vanligaste metoderna är att invertera, logaritmera och ta kvadratroten.

5 Analys av data

Innan man utför kanonisk korrelationsanalys är det lämpligt att först undersöka datasetet. Det man vill titta närmare på är bl a vad variablerna har för fördelning, om antagandena för metoden är uppfyllda enligt föregående kapitel, hur komplett datasetet är, om det saknas värden och om det finns extrema värden som skiljer sig avsevärt från mängden.

5.1 Labparametrar

Eftersom inte alla variabler är normalfördelade, börjar man med att transformera de så att de blir normalfördelade. CRP, LPK, Fibrinogen, Tg, LDL, Hb och EVF logaritmerades medan Kolesterol och HDL inverterades.

5.1.1 Saknade värden och outliers

Det saknas ett värde på CRP, Kolesterol och Fibrinogen som har orsakats på grund av slarv. För LDL och HDL saknas nästan hälften av värdena. Detta på grund av att man hade missat att ta dessa prover på vissa patienterna eftersom man inte hade fått instruktioner om det. Vi antar att bortfallen är slumpmässiga.

Ett sätt att hantera dataset med saknade värden är att inkludera bara de personer i analysen som har alla värden. Nackdelen är att datasetet blir mindre. I vårt fall skulle datasetet minskas enormt eftersom nästan hälften av LDL och HDL värdena saknades. Ett annat sätt är att ta bort de variabler som det saknas för många värden på. Då är det viktigt att man inte tar bort en variabel som är viktig för modellen. I vårt fall var LDL och HDL av stor betydelse och vi valde att ha kvar de i undersökningen. De två nyssnämnda metoderna ersätter inte saknade värden utan gör sig snarare av med dem.

De följande två metoder skattar saknade värden och ersätter dem.

Medelvärdessubstitution: man ersätter saknade värden för en variabel med medelvärdet av de värden som finns för variabeln.

Regressionsimputation: man predikterar saknade värden med regressionsanalys. Variabeln med de saknade värden blir responsvariabel och de övriga används som regressorer.

CRP-, kolesterol- och fibrinogenvärdena ersätts med medelvärdessubstitution. Anledningen till att de inte skattas med regressionsimputation är att det saknades för många värden på HDL och LDL och man skulle då antingen ta bort HDL och LDL eller så skulle man ha med dem men skattningen skulle göras på ca hälften av patientdata. Dessutom var det bara ett värde som saknades på varje variabel. Informationen är tillräcklig för att använda medelvärdessubstitution. Däremot ersattes saknade värden för LDL och HDL med regressionsimputation. Det fattades för många värden för att kunna använda medelvärdessubstitution. Man kan inte ersätta hälften av värdena med medelvärdet. Variansen skulle bli väldigt liten och missvisande.

Vidare bör man även titta på outliers. Genom att rita histogram eller Box-plottar kan man få en överskådlig bild över sina outliers. Även här kan man välja att ta bort personer med mycket outliers. I vårt fall skulle datasetet bli ännu mindre och därför behölls outliers-värden. Dessutom kan dessa patienters värde vara av stor betydelse för resultatet.

5.1.2 Test av antaganden

Nu när datasetet är helt komplett görs normalitetstest över variablerna igen för att se fördelningen. Resultaten för Shapiro-Wilks test av hypoteserna att variablerna är normalfördelade ges i Tabell 1.

Labparameter	W-statistika	P-värde
CRP	0,975	0,313
Kolesterol	0,971	0,219
LPK	0,982	0,596
Fibrinogen	0,973	0,272
Tg	0,965	0,121
HDL	0,961	0,080
LDL	0,972	0,224
Hb	0,968	0,164
EVF	0,972	0,239

Tabell 1: Shapiro-Wilks normalitetstest för labparametrar.

Vi kan inte förkasta nollhypotesen att variablerna är normalfördelade på 5 % - nivån.

Korrelation mellan variablerna anges i Tabell 2.

	CRP	Kolesterol	LPK	Fibrinogen	TG	HDL	LDL	Hb	EVF
CRP	1	0,033	0,399	0,618	0,079	0,309	0,170	-0,246	-0,190
Kolesterol	0,033	1	0,043	0,079	-0,101	-0,049	-0,852	-0,152	-0,160
LPK	0,399	0,043	1	0,545	0,181	0,464	0,004	-0,321	-0,265
Fibrinogen	0,618	0,079	0,545	1	-0,073	0,003	0,054	-0,231	-0,231
TG	0,079	-0,101	0,181	-0,073	1	0,529	-0,047	0,110	0,123
HDL	0,309	-0,049	0,464	0,003	0,529	1	0,097	-0,005	0,079
LDL	0,170	-0,852	0,004	0,054	-0,047	0,097	1	0,181	0,193
Hb	-0,246	-0,152	-0,321	-0,231	0,110	-0,005	0,181	1	0,964
EVF	-0,190	-0,160	-0,265	-0,231	0,123	0,079	0,193	0,964	1

Tabell 2: Korrelation mellan labparametrar.

Vi ser att Hb och EVF har den starkaste korrelationen 0,964 och Kolesterol och LDL har den näst starkaste korrelationen - 0,852. Notera att den negativa korrelationen mellan kolesterol och LDL beror på att kolesterol har inverterats.

Det som är kvar att undersöka är linjaritet och multikolinjaritet.

För att undersöka linjaritet tittar vi på en scatterplotsmatris över transformerade data (se Appendix). Vi ser att inga variabler avviker från linjaritet.

Kolinjaritet/multikolinjaritet undersöks genom att titta på toleransen och variansinflationen i Tabell 3.

Labparametrar	Tolerans	Variansinflation
CRP	0,428	2,336
Kolesterol	0,201	4,966
LPK	0,425	2,351
Fibrinogen	0,374	2,677
Tg	0,579	1,728
HDL	0,409	2,443
LDL	0,188	5,328
Hb	0,056	17,841
EVF	0,059	17,085

Tabell 3: Tolerans och variansinflation för labparametrar.

Tabellen visar att Hb och EVF har tolerans nära noll och variansinflation över 10. Detta visar att nästan kolinjaritet (eftersom korrelationskoefficienten är 0,964 och inte 1) råder bland dessa två. En av variablerna Hb eller EVF borde därför tas bort från analysen. (Notera att Kolesterol och LDL har de högsta variansinflationerna efter Hb och EVF). Efter att ha diskuterat med min handledare på CMM valde vi att ta bort Hb från modellen detta för att Hb egentligen inte är någon medicinsk relevant parameter för denna undersökning utan bara ett rutinvärde som mäts hos alla patienter. Resultat efter att ha tagit bort Hb anges i Tabell 4.

Labparametrar	Tolerans	Variansinflation
CRP	0,453	2,208
Kolesterol	0,205	4,889
LPK	0,437	2,288
Fibrinogen	0,385	2,599
Tg	0,600	1,666
HDL	0,421	2,377
LDL	0,191	5,228
EVF	0,799	1,252

Tabell 4: Tolerans och variansinflation efter att tagit bort Hb från modellen.

Nu är alla variansinflationer under 10 och Hb har tagits bort från modellen.

Nu är alla antaganden uppfyllda för labparametrarna.

5.2 Gener

Även här börjar man med att transformera data eftersom variablerna inte är normalfördelade. För att få alla variabler att bli normalfördelade görs följande transformationer:
CD1a, CD1c och Thromboxan togs kvadratroten ur och de övriga dvs CD1b, CD1d, CD1e och vα24 logaritmerades.

5.2.1 Test av antaganden

Resultaten av Shapiro-Wilks test av hypoteserna att variablerna är normalfördelade anges i Tabell 5.

Gener	W-Statistika	P-värde
CD1a	0,961	0,078
CD1b	0,983	0,648
CD1c	0,974	0,292
CD1d	0,963	0,094
CD1e	0,978	0,431
Thromboxan	0,957	0,048
Vα24	0,957	0,052

Tabell 5: Shapiro-Wilks normalitetstest.

P-värdena här är mindre än p-värden för labparametrar vilket visar att generna är, slarvigt uttryckt, mindre normalfördelade än labparametrarna. Notera att ett p-värde är 0,048 och alltså mindre än 0,05 men vi utgår ändå från att data kan betraktas som normalfördelade.

Korrelation mellan generna anges i Tabell 6.

	CD1a	CD1b	CD1c	CD1d	CD1e	Thromboxan	Vα24
CD1a	1	0,629	0,301	0,202	0,338	0,319	0,028
CD1b	0,629	1	0,448	0,331	0,664	0,424	0,234
CD1c	0,301	0,448	1	0,477	0,083	0,362	0,008
CD1d	0,202	0,331	0,477	1	0,361	0,642	-0,121
CD1e	0,338	0,664	0,083	0,361	1	0,281	0,140
Thromboxan	0,319	0,424	0,362	0,642	0,281	1	0,084
Vα24	0,028	0,234	0,008	-0,121	0,140	0,084	1

Tabell 6: Korrelation mellan generna.

Vi ser att CD1b och CD1e har den högsta korrelationen 0,664. Den näst högsta korrelationen 0,642 är mellan Thromboxan och CD1d.

I detta dataset saknades inga värden.

För att undersöka linjaritet tittar vi på scatterplotmatrisen (Se Appendix). Inga variabler avviker från linjaritet.

Test för kolinjaritet görs på samma sätt som för labparametrar dvs toleransen och variansinflationen för varje variabel beräknas och resultatet anges i Tabell 7.

Gener	Tolerans	Variansinflation
CD1a	0,570	1,754
CD1b	0,258	3,874
CD1c	0,540	1,853
CD1d	0,410	2,436
CD1e	0,412	2,429
Thromboxan	0,504	1,984
Vα24	0,855	1,170

Tabell 7: Tolerans och variansinflation för gener.

Här är alla variansinflationer under 10.

5.2.2 Outliers

På scatterplotmatrisen (Se Appendix) ser vi på plottarna för CD1c att det är en mindre grupp bestående av sju patienter som skiljer sig från de övriga. Dessa patienter är numrerade 284, 331 och 334-338 och är alla män. Detta bör undersökas närmare. Avsikten är dock inte att utesluta vissa patienter ur studien utan att uppmärksamma CMM på eventuella mönster. Därför delas datasetet upp i två grupper. En grupp bestående av dessa 7 patienter (grupp 2) och en grupp med resten av patienter dvs 47 patienter (grupp 1). Medelvärdet för alla labparametrar för dessa två grupper beräknas för att se om det förekommer någon skillnad. Resultat anges i Tabell 8.

Labparametrar	Medelvärde (Grupp 1)	Std.avv. (Grupp 1)	Medelvärde (Grupp 2)	Std.avv. (Grupp 2)	Differensen $\bar{x}_1 - \bar{x}_2$	Standardfel (Differens)
CRP	0,941	1,280	0,539	1,171	0,402	0,480
Kolesterol	0,223	0,041	0,255	0,047	-0,032	0,019
LPK	1,925	0,316	1,976	0,168	-0,051	0,078
Fibrinogen	1,294	0,225	1,329	0,241	-0,035	0,097
Tg	0,452	0,497	0,923	0,576	-0,471	0,229
HDL	0,919	0,173	0,937	0,068	-0,018	0,036
LDL	0,992	0,309	0,743	0,276	0,249	0,114
Hb	4,899	0,106	4,914	0,152	-0,015	0,059
EVF	3,659	0,102	3,657	0,157	0,002	0,061

Tabell 8: Medelvärde och standardavvikelse för grupp 1 och 2.

De variabler som har stor differens är CRP, Tg och LDL. Vad gäller CRP, är standardavvikelsen stor och det gör resultatet osäkert. För LDL är skillnaden inte så stor men

det kan ändå förklaras med att patient 284 och 334 som tillhör grupp 2 har låga LDL-värden 0,47 respektive 0,31 vilket kan jämföras med medelvärdet för LDL som är 0,96. För Tg kan anledningen vara att patient 335 som tillhör grupp 2 har ett extremt högt Tg-värde (2,2) vilket kan jämföras med medelvärdet för Tg som är 0,51 och det kan vara anledningen till att medelvärdet för grupp 2 är större. Grupp 2:s värden är annars inte utmärkande när det gäller övriga labparametrar. Vi ser inga tydliga mönster i analysen.

5.3 Korrelation mellan labparametrar och gener

Vi tittar på korrelationen mellan labparametrar och gener för att se om det redan nu förekommer några höga korrelationer. Det kan finnas höga parvisa korrelationer som man kan upptäcka innan man gör kanonisk korrelationsanalys. Resultat anges i Tabell 9.

	CD1a	CD1b	CD1c	CD1d	CD1e	Thromboxan	Vα24
CRP	0,091	0,281	0,147	0,230	0,113	0,165	0,017
Kolesterol	-0,176	-0,071	-0,217	-0,020	0,094	-0,042	0,091
LPK	0,070	0,174	-0,026	-0,055	0,059	0,056	0,173
Fibrinogen	-0,069	0,153	-0,061	0,048	0,113	0,039	0,068
Tg	0,125	-0,011	-0,195	0,098	0,143	-0,052	-0,085
HDL	0,095	0,132	0,104	0,182	0,035	0,038	-0,103
LDL	0,157	0,094	0,233	0,102	-0,131	0,107	-0,101
Hb	0,010	-0,196	0,051	0,208	-0,193	0,229	-0,026
EVF	0,054	-0,178	0,121	0,223	-0,217	0,273	-0,039

Tabell 9: Korrelation mellan labparametrar och gener.

Vi ser att den högsta korrelationen 0,281 är mellan CRP och CD1b vilket är lågt.

Nu är båda dataseten fullständiga och de fyller alla antaganden. Nu kan man tillämpa metoden på data.

5.4 Resultat från kanonisk korrelationsanalys

Vi har sju kanoniska funktioner. I Tabell 10 anges den kanoniska korrelationen för dessa sju funktioner och p-värde för F-test av hypoteserna att kanoniska korrelationen och alla efterföljande kanoniska korrelationer är noll.

Kanoniska funktioner	Kanonisk korrelation	Kanoniska rötter	P-Värde
1	0,596	0,355	0,602
2	0,506	0,256	0,838
3	0,447	0,200	0,922
4	0,329	0,108	0,973
5	0,275	0,076	0,971
6	0,146	0,021	0,984
7	0,041	0,002	0,962

Tabell 10: Resultat från kanonisk korrelation.

Vi ser att den högsta korrelationen är ca 0,596 vilket inte är starkt. Dessutom är inget test signifikant. Slutsatsen blir att det inte finns något uppenbart samband mellan generna och labparametrarna. Dessutom är den kvadrerade korrelationen, som ger en skattning av andelen delad varians mellan två kanoniska variabler, väldigt låg (0,355).

I Tabell 11 anges de standardiserade kanoniska koefficienterna för den första kanoniska funktionen.

Gener och labparametrar	Kanoniska vikter
CD1a	0,304
CD1b	-0,476
CD1c	0,571
CD1d	-0,286
CD1e	-0,426
Thromboxan	0,747
Vα24	0,232
CRP	0,504
Kolesterol	-0,395
LPK	0,663
Fibrinogen	-0,727
Tg	-0,573
HDL	-0,289
LDL	-0,179
EVF	0,830

Tabell 11: Standardiserade kanoniska koefficienter.

De variabler som har de högsta vikterna är Thromboxan, Fibrinogen och EVF men eftersom testet för första kanoniska funktionen inte är signifikant (se Tabell 10) skall man tolka det med viss försiktighet.

5.5 Analys av män och kvinnor som två separata grupper

Eftersom inga samband kunde påvisas mellan labparametrarna och gener kan man göra två olika analyser för män och kvinnor. Detta för att kanske kunna påvisa samband för män eller kvinnor. Av dessa 54 patienter är 17 kvinnor och 37 män. Vi delar in datasetet i män och kvinnor och gör om beräkningar först för kvinnor och sedan för män. Normalitetstestet visar att alla variabler (labparametrar och gener), både för män och kvinnor, är fortfarande normalfördelade. Vi börjar med att räkna ut medelvärde och standardavvikelse för män respektive kvinnor för alla variabler. Notera att dessa beräkningar görs på transformerade data och inte originaldata. Resultat anges i Tabell 12.

Labparametrar	Medelvärde (Män)	Std.avv. (Män)	Medelvärde (Kvinnor)	Std.avv. (Kvinnor)	Differensen $\bar{x}_m - \bar{x}_k$	Standardfel (Differens)
CRP	0,875	1,310	0,919	1,193	-0,044	0,361
Kolesterol	0,232	0,046	0,219	0,034	0,013	0,011
LPK	1,884	0,270	2,036	0,337	-0,152	0,093
Fibrinogen	1,247	0,213	1,412	0,213	-0,165	0,062
Tg	0,589	0,558	0,347	0,419	0,242	0,137
HDL	0,953	0,153	0,851	0,168	0,102	0,048
LDL	0,946	0,335	0,992	0,268	-0,046	0,085
Hb	4,919	0,115	4,861	0,093	0,058	0,029
EVF	3,679	0,113	3,615	0,087	0,064	0,028
CD1a	3,213	1,560	3,635	2,124	-0,422	0,575
CD1b	1,730	0,945	1,954	1,006	-0,224	0,289
CD1c	2,166	0,973	2,131	0,550	0,035	0,208
CD1d	0,722	0,611	0,467	0,676	0,255	0,192
CD1e	2,109	1,084	1,814	0,736	0,295	0,252
Thromboxan	2,648	0,702	2,391	0,451	0,257	0,159
Vα24	1,774	1,866	2,046	1,854	-0,272	0,544

Tabell 12: Medelvärde och standardavvikelse för alla variabler för män och kvinnor.

Vi ser att det förekommer skillnader mellan männens och kvinnornas värden men att få av dessa är signifikanta vilket inte är förvånande.

5.5.1 Analys av kvinnor

Vi börjar med att titta på korrelationen mellan labparametrar, mellan gener och slutligen mellan labparametrar och gener. Resultat anges i Tabellerna 13-15.

	CRP	Kolesterol	LPK	Fibrinogen	Tg	HDL	LDL	Hb	EVF
CRP	1	-0,164	0,444	0,730	0,322	0,451	0,290	0,127	0,158
Kolesterol	-0,164	1	0,125	0,070	-0,045	-0,068	-0,951	-0,090	-0,047
LPK	0,444	0,125	1	0,432	0,400	0,776	-0,051	-0,235	-0,229
Fibrinogen	0,730	0,070	0,432	1	0,208	0,140	-0,082	-0,091	-0,038
Tg	0,322	-0,045	0,400	0,208	1	0,626	0,108	0,115	0,004
HDL	0,451	-0,068	0,776	0,140	0,626	1	0,256	0,013	0,008
LDL	0,290	-0,951	-0,051	-0,082	0,108	0,256	1	0,171	0,122
Hb	0,127	-0,090	-0,235	-0,091	0,115	0,013	0,171	1	0,960
EVF	0,158	-0,047	-0,229	-0,038	0,004	0,008	0,122	0,960	1

Tabell 13: Korrelation mellan labparametrar.

Hb och EVF har fortfarande lika stark korrelation men korrelationen för LDL och Kolesterol är starkare, -0,951 jämfört med -0,852 (Se Tabell 2).

	CD1a	CD1b	CD1c	CD1d	CD1e	Thromboxan	Vα24
CD1a	1	0,808	0,638	0,198	0,422	0,260	0,080
CD1b	0,808	1	0,753	0,532	0,724	0,586	0,016
CD1c	0,638	0,753	1	0,747	0,658	0,424	-0,003
CD1d	0,198	0,532	0,747	1	0,501	0,389	-0,260
CD1e	0,422	0,724	0,658	0,501	1	0,480	-0,005
Thromboxan	0,260	0,586	0,424	0,389	0,480	1	0,091
Vα24	0,080	0,016	-0,003	-0,260	-0,005	0,091	1

Tabell 14: Korrelation mellan gener.

Den starkaste korrelationen 0,808 är mellan CD1a och CD1b. Den näst starkaste korrelationen finns mellan CD1c och CD1d som förut hade en svag korrelation (Se Tabell 6).

	CRP	Kolesterol	LPK	Fibrinogen	Tg	HDL	LDL	Hb	EVF
CD1a	0,227	-0,227	-0,089	-0,124	0,261	0,210	0,279	0,152	0,199
CD1b	0,504	-0,348	0,057	0,099	0,241	0,271	0,440	0,087	0,100
CD1c	0,224	-0,473	-0,156	-0,136	0,083	0,146	0,561	0,144	0,111
CD1d	0,339	-0,292	-0,151	-0,010	0,027	0,127	0,466	0,239	0,196
CD1e	0,249	-0,234	0,087	0,063	0,157	0,139	0,274	-0,023	-0,077
Thromboxan	0,222	-0,459	-0,058	0,216	0,194	-0,097	0,387	0,163	0,115
Vα24	-0,031	-0,225	0,155	-0,051	-0,163	0,115	0,198	0,320	0,309

Tabell 15: Korrelation mellan labparametrar och gener.

Den starkaste korrelationen är fortfarande mellan CRP och CD1b men den är högre nu än tidigare (Se Tabell 9).

Eftersom korrelationen mellan variablerna är något högre än tidigare testar vi kolinjaritet för att se om (nästan) kolinjaritet/multikolinjaritet förekommer bland fler variabler än Hb och EVF. För resultat se Tabell 16.

Labparametrar	Tolerans	Variansinflation
CRP	0,174	5,732
Kolesterol	0,013	79,507
LPK	0,050	19,921
Fibrinogen	0,186	5,379
Tg	0,102	9,788
HDL	0,023	44,216
LDL	0,009	110,693
Hb	0,020	50,424
EVF	0,020	50,527

Tabell 16: Kolinjaritetstest för labparametrar.

Här är det fler variabler som har variansinflation över 10. Förutom Hb tar vi även bort LDL, som har den högsta variansinflationen, och testar kolinjaritet på nytt. Resultat anges i Tabell 17.

Labparametrar	Tolerans	Variansinflation
CRP	0,250	3,994
Kolesterol	0,843	1,186
LPK	0,178	5,607
Fibrinogen	0,229	4,370
Tg	0,502	1,994
HDL	0,135	7,433
EVF	0,782	1,278

Tabell 17: Kolinjaritetstest för labparametrar efter att ha tagit bort Hb och LDL.

Nu är alla variansinflationer för labparametrar under 10 och vi testar samma sak för gener. För resultat se Tabell 18.

Gener	Tolerans	Variansinflation
CD1a	0,109	9,207
CD1b	0,077	13,059
CD1c	0,136	7,360
CD1d	0,178	5,633
CD1e	0,310	3,226
Thromboxan	0,489	2,044
Vα24	0,800	1,251

Tabell 18: Kolinjaritetstest för gener.

Vi ser att nästan kolinjaritet förekommer även bland gener och att CD1b ska tas bort från modellen. Vi gör om kolinjaritetstestet efter att ha tagit bort CD1b. Resultat anges i Tabell 19.

Gener	Tolerans	Variansinflation
CD1a	0,407	2,454
CD1c	0,157	6,377
CD1d	0,253	3,946
CD1e	0,517	1,935
Thromboxan	0,712	1,404
Vα24	0,809	1,237

Tabell 19: Kolinjaritetstest för gener efter att ha tagit bort CD1b.

Nu när alla variansinflationer är under 10 kan vi utföra kanonisk korrelationsanalys på data. Resultat anges i Tabell 20. Notera att LDL, Hb och CD1b har tagits bort från modellen.

Kanoniska funktioner	Kanoniska korrelationer	Kanoniska rötter	P-värde
1	0,876	0,767	0,977
2	0,684	0,468	0,995
3	0,641	0,411	0,987
4	0,513	0,263	0,982
5	0,295	0,087	0,968
6	0,245	0,060	0,757

Tabell 20: Resultat från kanonisk korrelationsanalys.

Den högsta korrelationen 0,876 är högre jämfört med innan indelningen (Se Tabell 10) men testet är fortfarande icke signifikant. Det finns inget signifikant samband mellan labparametrarna och generna och resultaten bör tolkas med försiktighet. Notera att beräkningarna har gjorts på ett litet dataset och det höga p-värdet 0,997 kan förklaras av detta. En utökning av datasetet skulle kunna ge ett annat resultat än det vi har erhållit.

Standardiserade kanoniska koefficienter för den första kanoniska funktionen anges i Tabell 21.

Gener och labparametrar	Kanoniska vikter
CD1a	0,891
CD1c	-0,877
CD1d	1,432
CD1e	-0,111
Thromboxan	-0,709
Vα24	0,371
CRP	0,717
Kolesterol	0,419
LPK	-1,077
Fibrinogen	-0,365
Tg	-0,553
HDL	1,323
EVF	0,098

Tabell 21: Kanoniska vikter för den första funktionen.

Vikterna är högre jämfört med tidigare (Se Tabell 11) och variabler med de högsta vikterna är CD1d, LPK och HDL alltså inga av de variabler som hade de högsta vikterna innan.

5.5.2 Analys av män

På samma sätt som vi gjorde för kvinnorna, börjar vi först med att titta på korrelationen mellan labparametrar, mellan gener och mellan labparametrar och gener. Resultat anges i Tabellerna 22-24.

	CRP	Kolesterol	LPK	Fibrinogen	Tg	HDL	LDL	Hb	EVF
CRP	1	0,097	0,395	0,621	0,013	0,276	0,131	-0,372	-0,303
Kolesterol	0,097	1	0,066	0,164	-0,160	-0,109	-0,828	-0,224	-0,253
LPK	0,395	0,066	1	0,554	0,190	0,454	0,004	-0,307	-0,215
Fibrinogen	0,621	0,164	0,554	1	-0,068	0,103	0,073	-0,187	-0,191
Tg	0,013	-0,160	0,190	-0,068	1	0,463	-0,070	0,046	0,085
HDL	0,276	-0,109	0,454	0,103	0,463	1	0,073	-0,121	-0,004
LDL	0,131	-0,828	0,004	0,073	-0,070	0,073	1	0,214	0,248
Hb	-0,372	-0,224	-0,307	-0,187	0,046	-0,121	0,214	1	0,964
EVF	-0,303	-0,253	-0,215	-0,191	0,085	-0,004	0,248	0,964	1

Tabell 22: Korrelation mellan labparametrar.

Korrelationerna har inte ändrats så mycket jämfört med innan indelning av datasetet (Se Tabell 2). Hb och EVF har fortfarande den starkaste korrelationen 0,964. Kolesterol och LDL har den näst starkaste korrelationen -0,828.

	CD1a	CD1b	CD1c	CD1d	CD1e	Thromboxan	Vα24
CD1a	1	0,514	0,222	0,252	0,359	0,412	-0,015
CD1b	0,514	1	0,392	0,271	0,697	0,430	0,328
CD1c	0,222	0,392	1	0,432	-0,017	0,356	0,013
CD1d	0,252	0,271	0,432	1	0,302	0,728	-0,037
CD1e	0,359	0,697	-0,017	0,302	1	0,219	0,201
Thromboxan	0,412	0,430	0,356	0,728	0,219	1	0,104
Vα24	-0,015	0,328	0,013	-0,037	0,201	0,104	1

Tabell 23: Korrelation mellan gener.

Den högsta korrelationen är mellan Thromboxan och CD1d som förut hade den näst starkaste korrelationen. Den näst starkaste korrelationen är mellan CD1b och CD1e som förut hade den starkaste korrelationen. Korrelationen mellan generna har alltså inte ändras så mycket jämfört med innan indelningen (Se Tabell 6).

	CRP	Kolesterol	LPK	Fibrinogen	Tg	HDL	LDL	Hb	EVF
CD1a	0,017	-0,141	0,145	-0,114	0,114	0,087	0,096	-0,013	0,042
CD1b	0,185	0,048	0,211	0,136	-0,067	0,122	-0,044	-0,278	-0,250
CD1c	0,133	-0,176	0,018	-0,040	-0,260	0,096	0,172	0,029	0,124
CD1d	0,193	0,040	0,076	0,187	0,073	0,140	-0,017	0,145	0,180
CD1e	0,081	0,146	0,105	0,210	0,107	-0,055	-0,219	-0,289	-0,313
Thromboxan	0,161	0,012	0,172	0,086	-0,160	0,006	0,062	0,200	0,262
Vα24	0,035	0,213	0,167	0,090	-0,042	-0,185	-0,215	-0,127	-0,134

Tabell 24: Korrelation mellan gener och labparametrar.

Här är korrelationerna fortfarande svaga. Den starkaste -0,289 är mellan Hb och Cd1e. Vi testar kolinjaritet för att se om resultaten är fortfarande samma som innan indelningen. Det borde de vara eftersom korrelationerna inte har ökat så mycket jämfört med tidigare. Resultat för kolinjaritetstest av labparametrar anges i Tabell 25.

Labparametrar	Tolerans	Variansinflation
CRP	0,429	2,333
Kolesterol	0,209	4,786
LPK	0,447	2,239
Fibrinogen	0,370	2,702
Tg	0,631	1,585
HDL	0,552	1,811
LDL	0,204	4,907
Hb	0,045	22,117
EVF	0,049	20,397

Tabell 25: Kolinjaritetstest för labparametrar.

Hb och EVF har variansinflationer över 10. Av samma anledning som tidigare tar vi bort Hb från modellen. Resultat för kolinjaritet efter att ha tagit bort Hb anges i Tabell 26.

Labparametrar	Tolerans	Variansinflation
CRP	0,484	2,066
Kolesterol	0,209	4,785
LPK	0,511	1,955
Fibrinogen	0,428	2,337
Tg	0,638	1,568
HDL	0,578	1,730
LDL	0,204	4,907
EVF	0,779	1,284

Tabell 26: Kolinjaritetstest för labparametrar efter att ha tagit bort Hb.

Vi testar samma sak för gener och anger resultatet i Tabell 27.

Gener	Tolerans	Variansinflation
CD1a	0,649	1,541
CD1b	0,210	4,763
CD1c	0,463	2,161
CD1d	0,287	3,481
CD1e	0,282	3,543
Thromboxan	0,313	3,193
Vα24	0,814	1,229

Tabell 27: Kolinjaritetstest för gener.

Här är alla variansinflationer under 10 och vi kan nu utföra kanonisk korrelationsanalys på data. Resultat anges i Tabell 28. Notera att Hb har tagits bort från modellen.

Kanoniska funktioner	Kanoniska korrelationer	Kanoniska rötter	P-värde
1	0,789	0,623	0,488
2	0,592	0,351	0,939
3	0,457	0,209	0,975
4	0,419	0,175	0,962
5	0,337	0,114	0,960
6	0,225	0,051	0,955
7	0,061	0,004	0,949

Tabell 28: Resultat från kanonisk korrelationsanalys.

Den högsta kanoniska korrelationen 0,789 är högre än innan indelningen (Se Tabell 10) men lägre än kvinnornas (Se Tabell 20). Testet är fortfarande icke signifikant och resultaten bör därför tolkas med försiktighet. Även här kan man inte påvisa samband mellan labparametrar och gener.

De standardiserade kanoniska koefficienterna för den första kanoniska funktionen anges i Tabell 29.

Gener och labparametrar	Kanoniska vikter
CD1a	0,132
CD1b	-0,744
CD1c	0,752
CD1d	-1,087
CD1e	0,100
Thromboxan	1,274
Vα24	0,086
CRP	0,536
Kolesterol	-0,303
LPK	0,730
Fibrinogen	-0,832
Tg	-0,686
HDL	-0,314
LDL	-0,158
EVF	0,662

Tabell 29: Kanoniska vikter för den första funktionen.

Här är vikterna högre än innan indelningen (Se Tabell 11) och de variabler som har de högsta vikterna är CD1d, Thromboxan och Fibrinogen. Notera att Thromboxan och Fibrinogen fanns bland de variabler med de högsta vikterna även innan indelningen av datasetet.

6 Sammanfattning

6.1 Kort sammanfattning av resultaten

Det går inte att påvisa något signifikant samband mellan de kliniska labparametrarna och generna. Detta kan man se på den låga kanoniska korrelationen.

Korrelationen mellan labparametrar och gener visade att inga starka samband förekom. 0,281 var den högsta parvisa korrelationen. Det kan då tyckas onödigt att göra kanonisk korrelationsanalys. Man kanske borde ha stannat då och dragit den slutsatsen att inga samband förekom. Men tanken med att utföra kanonisk korrelationsanalys är att se om det kunde finnas starka multipla samband snarare än bara parvisa samband. Med multipla samband menas om några labparametrar tillsammans har samband med några gener. Tyvärr var så inte fallet.

Förhoppningen var att de två separata analyserna på män och kvinnor skulle ge andra resultat. Det visade sig att indelning av kön inte heller påverkade resultatet avsevärt. Det finns inget signifikant samband mellan de kliniska labparametrarna och generna varken för män eller kvinnor. Däremot var kvinnornas resultat annorlunda än männens. För kvinnor var sambandet mellan Kolesterol och LDL väldigt starkt. Man får inte glömma att antalet kvinnor var endast 17 stycken. Flera kvinnliga patienter skulle kanske kunna påvisa signifikanta samband mellan labparametrar och gener.

Däremot är ett resultat säkert. Ett starkt samband förekommer mellan Hb och EVF (0,964). Detta gällde alla analyser. Ett viktigt resultat är om man har ena variabeln behöver man inte ha den andra. De förklarar varandra mycket bra. Man kan här spara tid och pengar genom att bara mäta den ena variabeln, förslagsvis den som är billigast eller enklast.

Man bör tänka på att det här är en relativt liten studie med bara 54 patienter. Om man utökade den skulle man kanske finna signifikanta samband. Dessutom får man inte glömma det stora bortfallet för HDL och LDL som kan ha påverkat resultatet något.

6.2 Diskussion

Hade tid funnits, skulle man kunnat lägga mer tid och arbete på extrema värden. Bäst vore om man kunde få helt nya testvärden på patienter med outliers för att försäkra sig att mätningarna är rätt gjorda. På befintligt data kan man annars med t ex Klusteranalys undersöka outliers lite närmare.

En av svårigheterna med medicinskt data är att datasetet kan vara mycket begränsat. Datasetet ökar när en person insjuknar och lämnar prover. Man får oftast nöja sig med ett mindre dataset och dra sina slutsatser från analys av detta. Värt att notera är att framtagning av liknande dataset som vi sett i analysen är en process som tar tid och kan kosta mycket pengar.

Vad man också önskar är en jämförelse mellan friska och sjuka patienter. Men det är svårt att få tag på friska personer som frivilligt går med på att prover tas från halspulsådern!

7 Referenslista

[1] Henrik Brändén (1997), Molekylär biologi

[2] William R. Dillon, Matthew Goldstein (1984), Multivariate Analysis, Methods And Applications

[3] Per Fernlund, Arne Hanson, Carl-Bertil Laurell, Bengt Lundh (1986), Klinisk kemi i praktisk medicin, 5:e upplagan

[4] Joseph F. Hair JR, Rolph E. Anderson, Ronald L. Tatham, William C. Black (1998), Multivariate Dataanalysis, 5:e upplaga

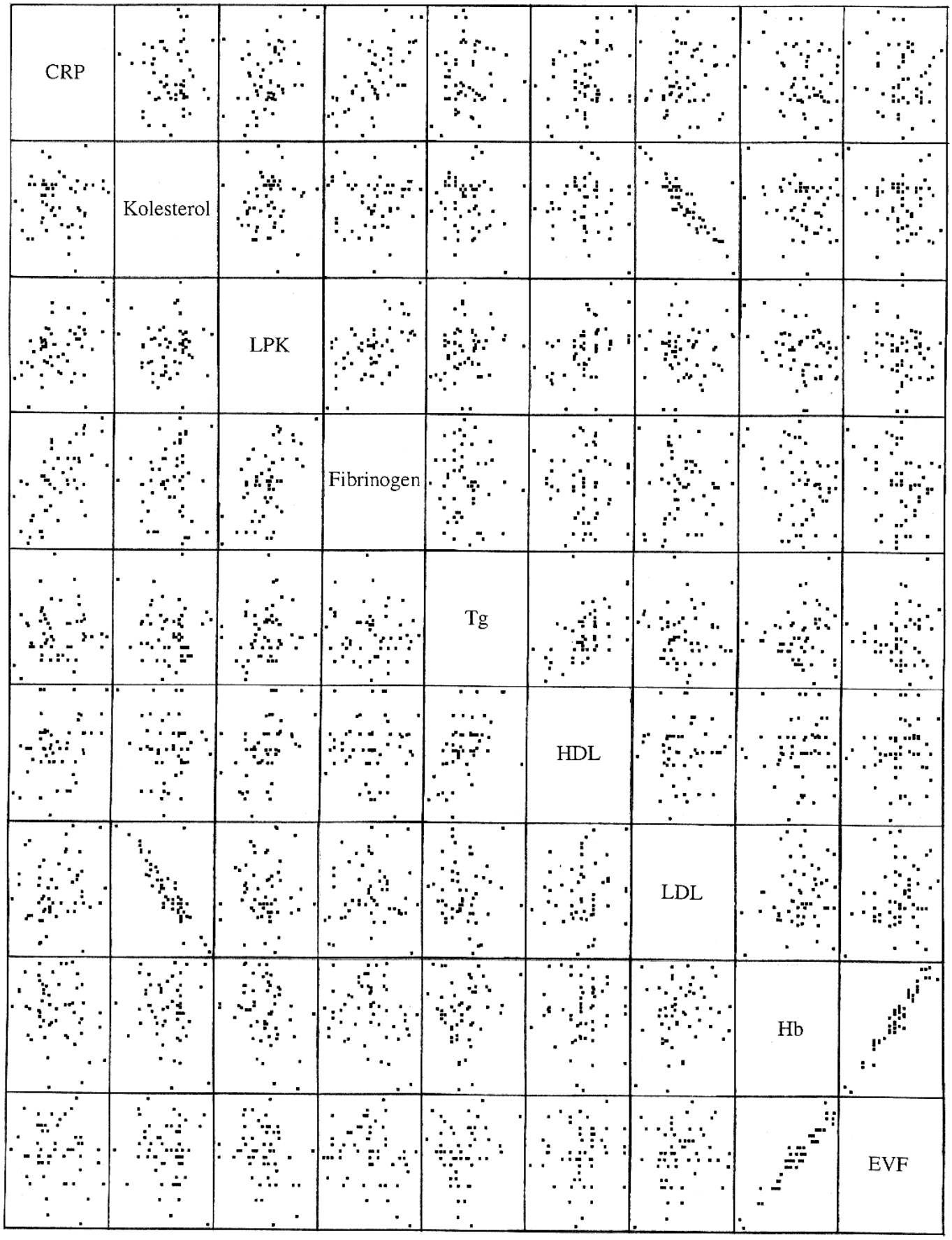
[5] Richard A. Johnson, Dean W. Wichern (1988), Applied Multivariate Statistical Analysis, 2:a upplagan

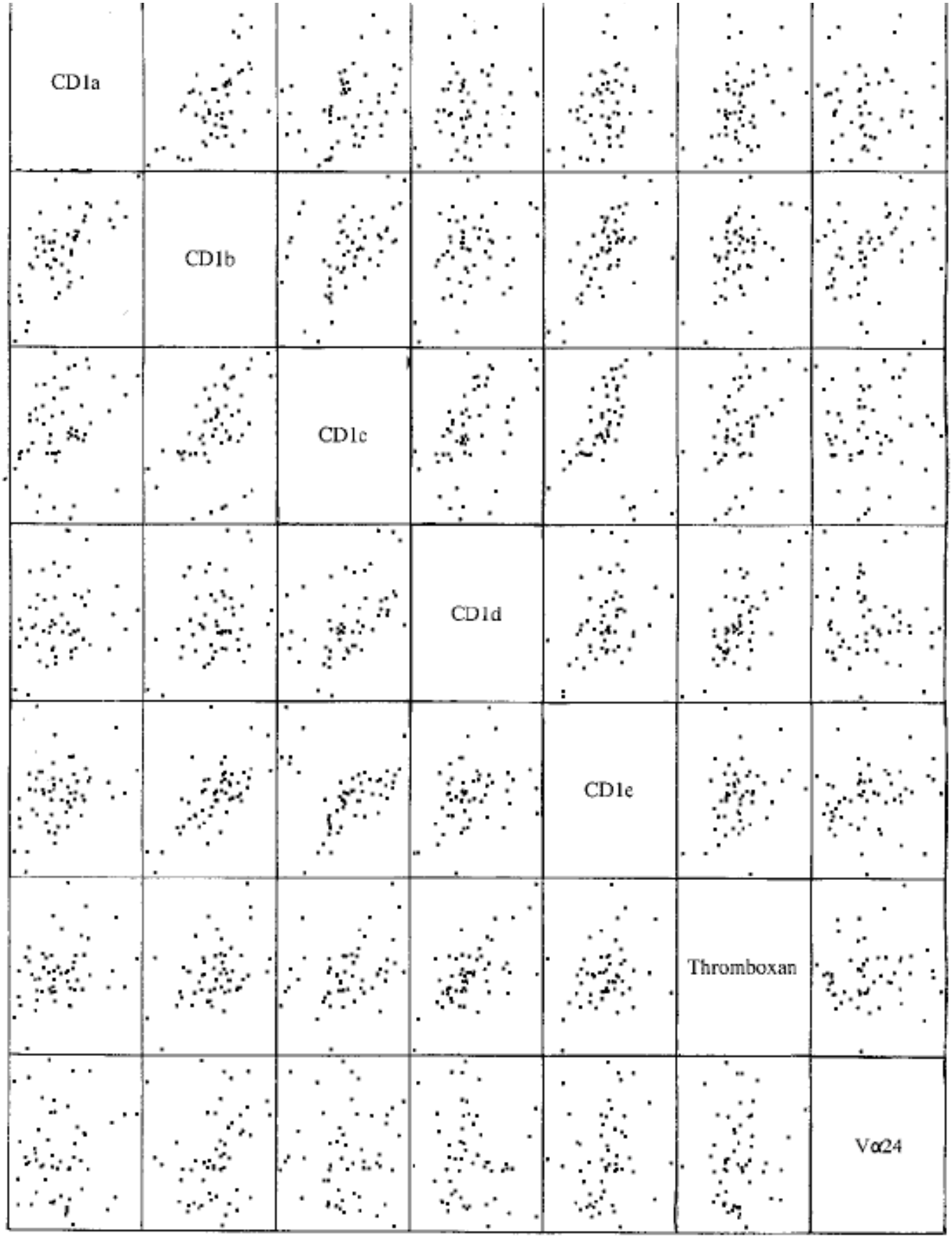
[6] Rolf Sundberg (2002) Collinearity. Encyclopedia of Environmetrics (ed. El-Shaarawi and Piegorsch) Volume 1, pp 365-366. John Wiley & Sons, Chichester.

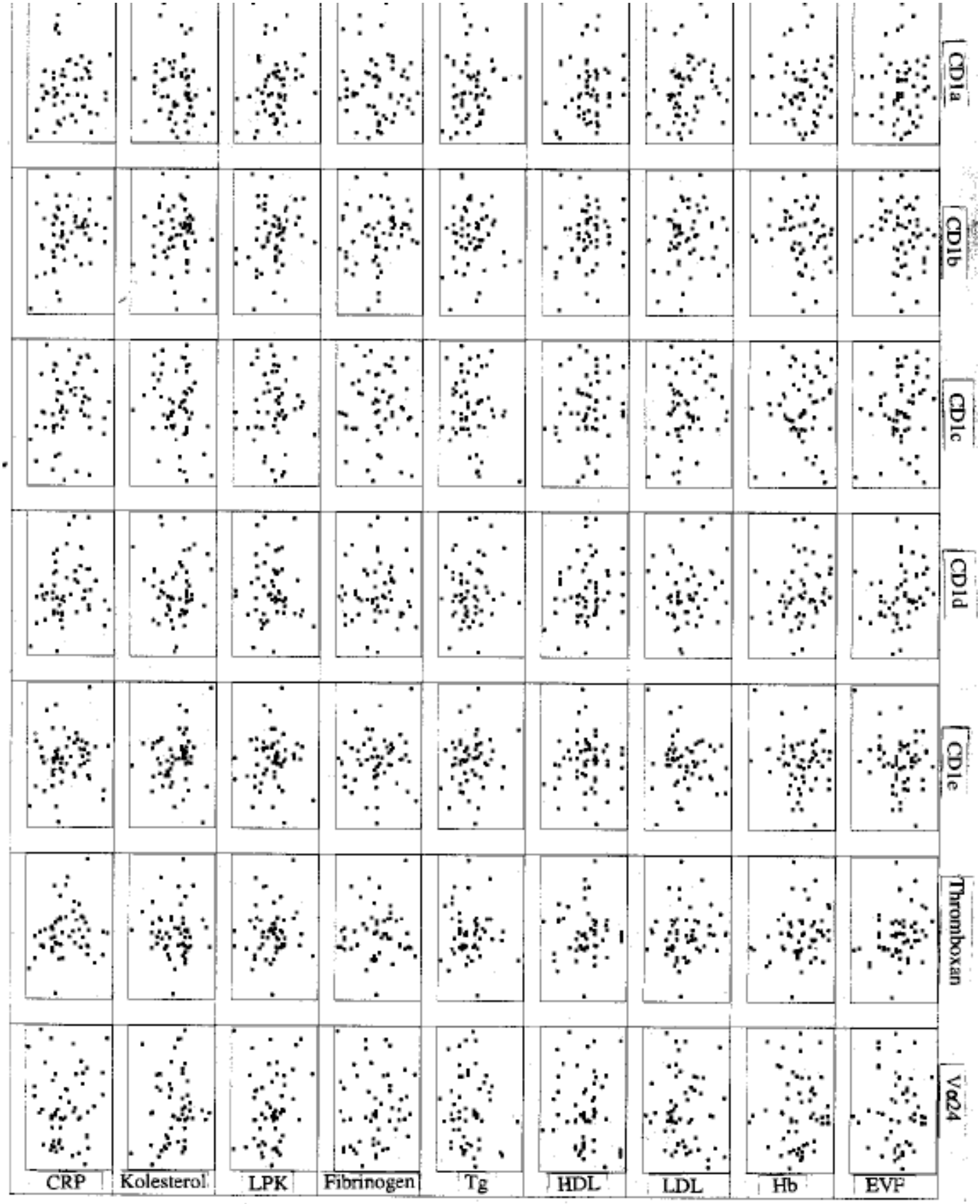
[7] SAS hjälpfunktion

[8] www.ne.se

Appendix


CRP
Kolesterol
LPK
Fibrinogen
Tg
HDL
LDL
Hb
EVF


CD1a
CD1b
CD1c
CD1d
CD1e
Thromboxan
Vα24


CD1a
CD1b
CD1c
CD1d
CD1e
Thromboxan
Vα24
CRP
Kolesterol
LPK
Fibrinogen
Tg
HDL
LDL
Hb
EVF