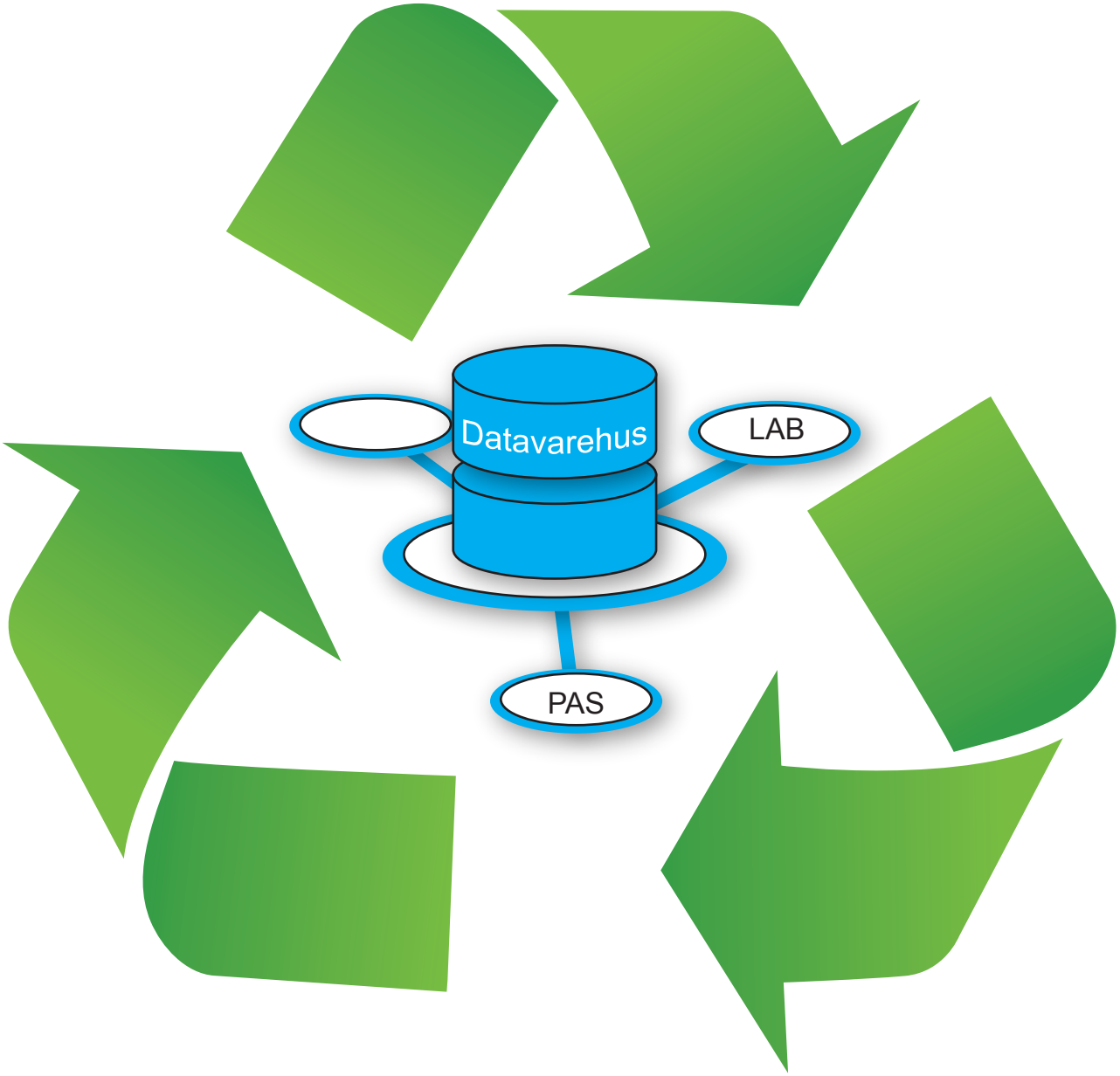
Mathias Abitz Boysen Julian Birkemose Nielsen



Identificering af kemiske biomarkører til detektering af kolorektal kræft ved hjælp af eksisterende data

Et projekt om anvendelse af et dimensionelt datavarehus i sundhedssektoren

Gruppe 13gr1073

Aalborg Universitet - Civilingeniøruddanelsen i Sundhedsteknologi - 4. Juni 2013

AALBORG UNIVERSITET

Institut for Sundhedsvidenskab & Teknologi
Frederik Bajers Vej 7D2
9220 Aalborg Ø, Danmark
Telefon (+45) 99 40 99 40
http://www.hst.aau.dk

Titel: Identificering af kemiske biomarkører til detektering af kolorektal kræft ved hjælp af eksisterende data
– Et projekt om anvendelse af et dimensionelt datavarehus i sundhedssektoren

Tema: Speciale projekt

Projektperiode:
Foråret 2013

Projektgruppe:
Gruppenummer: 13gr1073

Gruppemedlemmer:
Mathias Abitz Boysen
Julian Birkemose Nielsen

Vejleder:
Pia Britt Elberg

Oplagstal: 5

Rapport sideantal: 78

Appendix sideantal: 19

Totale sideantal: 110

Projekt gennemført d.: 04-06-2013

Synopsis:

I Danmark diagnosticeres ca. 4.500 med kolorektal kræft (KRC) hvert år. Den 5-årige overlevelse er observeret til ca. 50%, da KRC ofte diagnosticeres sent. Derfor oprettes et screeningsprogram i Danmark i 2014. Dog kritiseres nuværende screeningsmetoder, hvorfor der er fokus på identificering af kemiske biomarkører. Der er endnu ikke identificeret brugbare biomarkører til detektering af KRC.
Region Nordjyllands datavarehus, der indeholder laboratorie- og patientadministrative data, kan give mulighed for data mining efter nye, potentielle kemiske biomarkører for KRC.
15 biokemiske analysetyper blev undersøgt med lineær regression, middelværdi og standardafvigelse vha. mønstergenkendelse i perioden, før KRC diagnosticeres hos patienterne. Lineær og kvadratisk diskriminant analyse blev brugt til at adskille 125 patienter med KRC fra 6.149 patienter uden kræft. Data blev normaliseret for at opnå uafhængighed af laboratoriernes normalværdier.
Den simpleste klassifikator ift. højeste sensitivitet og specificitet var en 3-dimensionel kvadratisk klassifikator, som udnyttede lineær regression af *P-Albumin*, middelværdi af *B-Thrombocytter* og *P-Bilirubiner*. Sensitiviteten var 92,0% og specificteten var 79,5%.
Dataanalysen er retrospektiv, så kompletheden og fordelingen af data i datavarehuset, er de største begrænsninger i projektet, da det nedsætter antallet af samples og kan skævride resultaterne.
Trods begrænsningerne har klassifikatoren højere sensitivitet men lavere specificitet end fækal okkult blodtest, hvilket betyder at klassifikatoren har potentiale til at detekterer en større procentdel af de patienter, der har KRC, end fækal okkult blodtest. Dermed har datavarehuset med mønstergenkendelse identificeret mulige kemiske biomarkører for KRC.

AALBORG UNIVERSITY

Department of Health Science & Technology
Frederik Bajers Vej 7D2
9220 Aalborg Ø,Denmark
Telephone (+45) 9940 9940
http://www.hst.aau.dk

Title: Identification of chemical biomarkers for detection of colorectal cancer using existing data
– A project about the use of a dimensional data warehouse in healthcare

Theme: Master's thesis

Project period:
Spring 2013

Project group:
Group number: 13gr1073

Group members:
Mathias Abitz Boysen
Julian Birkemose Nielsen

Supervisor
Pia Britt Elberg

No of copies: 5

No. of pages: 78

No of appendix pages: 19

Total no. of pages: 110

Completed: 04-06-2013

Synopsis:

Every year approx. 4.500 people are diagnosed with colorectal cancer (CRC) in Denmark. The 5-year survival rate has been observed to be approx. 50% as CRC is often diagnosed at a late stage. In 2014 a screening program will be established in Denmark. However, current screening methods are being criticized, making chemical biomarkers the center of attention. At present no useful biomarkers for detecting CRC have been identified.
Data mining in *The North Denmark Region's* data warehouse, that contains laboratory and patient administrative data, should allow for detection of new, possible chemical biomarkers for CRC.
Using pattern recognition, 15 biochemical tests were investigated with linear regression, mean and standard deviation, in the period before patients were diagnosed with CRC. Linear and quadratic discriminant analyses were used to separate 125 patients with CRC from 6,149 patients without cancer. Data were normalized to obtain independence of the laboratories' normal values.
The simplest classifier in relation to the highest sensitivity and specificity was a 3-dimensional quadratic classifier based on the linear regression of *P-Albumin* including the mean value of *B-Thrombocytes* and *P-Bilirubin*. Sensitivity was 92% and specificity was 79.5%.
Completeness and distribution of the data in the data warehouse are major limitations in this project, due to the retrospective nature, as it reduces the number of samples and may skew the results.
Despite the limitations, the classifier has a higher sensitivity but a lower specificity compared to fecal occult blood test, meaning that the classifier could detect a greater percentage of patients with CRC compared to the fecal occult blood tests. Thus, possible chemical biomarkers for CRC have been identified with the data warehouse through pattern recognition.

Abstract

Looking at the prevalence of cancers affecting both sexes, colon and rectum cancer is one of the most common. Colon and rectum cancer is also known as colorectal cancer (CRC). CRC is the third most frequently diagnosed cancer world wide and it is one of the more aggressive cancers, with an expected 5-year survival rate of about 50%. This is despite the fact that the 5-year survival rate can be as high as 90%, if CRC is detected at an early stage. Every year approx. 4.500 people are diagnosed with CRC in Denmark. The issue has been acknowledged, because a new screening program for CRC will be established in Denmark in 2014. However, current screening methods are being criticized for varying sensitivity, lowering the participation rates and being expensive. Thereby making chemical biomarkers the center of attention, since they have the potential to yield fast and minimally invasive detection of CRC. At present no useful chemical biomarkers for detecting CRC have been identified, but chemical biomarkers remain the area of interest because many of the chemical biomarkers have been linked to the carcinogenesis of CRC. It has been hypothesized that one biomarker alone cannot describe the complexity of carcinogenesis, which is why combining biomarkers has been suggested. This will guide the data analysis in this project.

The healthcare sector has generated enormous amounts of data that has the potential to support research, such as identifying new chemical biomarkers for CRC. *The North Denmark Region's* data warehouse makes it possible to test this theory, because it has recently been expanded to contain both laboratory and patient administrative data.

Before the data analysis could be performed there were issues regarding secondary use of data that had to be addressed. First off data had to be untangled from their context before using them for secondary purposes. Initially this was addressed through classification of the data. The danish healthcare sector is already classifying both data from the patient administrative system and data from the laboratory information system with SKS codes and NPU codes, respectively. Another context issue that could not be solved with classification was the dependence of the laboratory data on the different normal values that each laboratory has. To obtain independence of the normal values of the laboratories and allow for comparison of data across different laboratories, laboratory data were normalized.
Secondly, the integration of data from the patient administrative system and laboratory data had to be addressed. Contrary to popular belief a data warehouse does not solve the integration issue, it only facilitates a solution. Integration of data is dependent on the context and purpose of the data analysis. In this project the test results from the laboratory had to be integrated with diagnosis codes from a year before the diagnosis to the time of the diagnosis.

As there are no chemical biomarkers for detecting CRC pattern recognition was utilized to investigate 15 biochemical tests with linear regression, mean and standard deviation, in a one

year period before the patients were diagnosed with CRC. Linear and quadratic discriminant analysis' were used to separate 125 patients with CRC from 6,149 patients without cancer. The results showed that the best performing classifier with the lowest complexity in relation to the highest sensitivity and specificity was a 3-dimensional quadratic classifier based on the linear regression of *P-Albumin* including the mean values of *B-Thrombocytes* and *P-Bilirubin*. This classifier yielded a sensitivity of 92% and specificity of 79.5%. Compared with the fecal occult blood test the sensitivity is about the same, but varying sensitivity has been reported, potentially making the sensitivity of the classifier higher. However, the specificity of the classifier is lower than that of FOBT.

Due to the retrospective nature of this project, the two major limitations were the completeness and distribution of the data in the data warehouse, as it may skew the results and it greatly reduces the number of samples available for analysis.

Despite the limitations, the classifier could have a higher sensitivity compared to fecal occult blood test, meaning that the classifier could detect a greater percentage of patients with CRC compared to the fecal occult blood tests. Thus, possible chemical biomarkers for CRC have been identified with the data warehouse through pattern recognition. However, because the project is retrospective the conclusion is not definitive. New analyses with more samples are necessary to validate the results. More data are also necessary to make sure that the results are not due to confounding factors. Lastly, prospective studies of the chemical biomarkers are necessary to obtain a stronger conclusion.

Forord

Dette speciale projekt er udarbejdet af projektgruppe 13gr1073 på fjerde og afsluttende semester af kandidatuddannelsen til Civilingeniør i Sundhedsteknologi. Specialet projektet er udarbejdet i perioden fra d. 1. februar 2013 til 4. juni 2013 ved Aalborg Universitet.

Dette speciale er et datavarehus anvendelsesprojekt, som følger op på et tidligere datavarehus udviklingsprojekt. Det tidligere projekt [1] omhandlede udviklingen af en datamodel for laboratoriedata og implementeringen af modellen i datavarehuset, som anvendes i dette projekt til at detektere kolorektal kræft.

Specialet projektets formål var, at undersøge hvilke muligheder Region Nordjyllands datavarehus giver i relation til klinisk forskning i forbindelse med detektering af kolorektal kræft. Datavarehuset i dette projekt er endnu ikke i drift i Region Nordjylland. Projektet omhandler genbrug af data til at løse en kliniske problemstilling. Dette projekt har en tværfaglig karakter. Metodisk er projektet et data mining projekt, som udnytter mønstergenkendelse, men klinisk informatik spiller også en stor rolle i at kunne forstå og overkomme de problemstillinger, som genbrug af data til sekundære formål introducerer.

Projektgruppen ønsker at rette en stor tak til Region Nordjylland for adgang til anonymiserede data. Derudover ønsker projektgruppen at takke Enversion for teknisk støtte i forbindelse med brugen af datavarehuset, *FaktaCenter*.
Der rettes en særlig tak til Torleif Trydal, Overlæge, Aalborg Universitetshospital og Simon Lykkeboe, Biokemiker, Aalborg Universitetshospital for at have bidraget med sparring og viden ift. projektets ide.

Aalborg Universitet, 4. juni 2013

Mathias Abitz Boysen

Julian Birkemose Nielsen

Indholdfortegnelse

Indledning 1

På verdensbasis er kræft årsagen til 21% af alle sygdomsrelaterede dødsfald, hvilket gør kræft til den næst hyppigste dødsårsag i verden, kun overgået af kardiovaskulære sygdomme. I 2008 døde 7,6 millioner mennesker af kræft på verdensbasis, og dette tal forventes at stige til 13 millioner i 2030 [2, s. 33-37]. I Danmark diagnosticeres der årligt op imod 32.000 danskere med kræft, og af disse er det kun ca. 50%, som lever 5 år efter, at de har fået diagnosen. Dette kan skyldes, at ved diagnosetidspunktet har 2/3 af alle kræfttilfældene spredt sig til andre dele af kroppen.

Ser man på incidensraterne, så er den hyppigste kræftform hos mænd, prostatakræft. Den hyppigste kræftform hos kvinder er til gengæld brystkræft. Ser man på kræftformer der rammer begge køn, så er kræft i tyk- og endetarmen, også kaldet kolorektal kræft (KRC), en af de hyppigste kræftformer med ca. 4.500 nye tilfælde hvert år i Danmark. KRC er den tredje hyppigst diagnosticerede kræftform i verden og den resulterer i det næsthøjeste antal dødsfald blandt kræfttilfælde [3]. KRC er en af de mere aggressive kræftformer, hvor kun ca. 50% har en forventet overlevelse på 5 år efter diagnosetidspunktet[4, s. 143][5]. Dette er på trods af, at den 5 årige overlevelsesrate kan være helt op imod 90%, såfremt at tilfældene opdages tidligt i sygdomsforløbet [6, 7, 8, 9][4, s. 143][10, s. 938][11, s. 16]. Dvs. at tidlig detektering af KRC er yderst interessant, for desto tidligere man kan finde og behandle kræfttilfælde, desto større er overlevelsesraten. Herunder ses de forskellige stadier for KRC, og deres indflydelse på den 5 årige overlevelse [12, s. 6-15][6]. Følgende stadieinddeling er fra TNM systemet, jvf. appendiks A.2 på side 81.

- **Stadie 1** - Kræften er ikke vokset igennem tarmvæggen, og der er ingen spredning til lymfeknuder eller andre metastaser. 90-100% overlevelse.
- **Stadie 2** - Kræften er vokset igennem tarmvæggen, og evt. ind i andre organer. Der er ingen spredning til lymfeknuder eller andre metastaser. 75-85% overlevelse.
- **Stadie 3** - Der er spredning til de regionale lymfeknuder. 30-40% overlevelse.
- **Stadie 4** - Der er fjernmetastaser. < 5% overlevelse.

Sandsynligheden, for at overleve KRC, falder kraftigt desto senere kræften diagnosticeres. Dette er en anerkendt problemstilling, som forsøges løst ved tidlig identificering af kræften igennem screeningsprogrammer. Screeningsprogrammer kan indeholde et udvalg af forskellige metoder, hvorved målet om tidlig diagnosticering og behandling kan opnåes [12, s. 6-15]. De mest brugt metoder til screening for KRC er fækal okkult blodtest, fækal DNA test, sigmoideoskopi,

koloskopi og virtuel koloskopi. I Danmark beskrives de nationale screeningsprogrammer for kræft i tre kræftplaner. Den første kræftplan blev publiceret i 2000, og er løbende blevet opdateret og sidenhen efterfulgt af kræftplan II og III i takt med, at nye metoder og forskning er blevet tilgængelig [13, 14, 15]. Disse kræftplaner har bevirket, at der i dag tilbydes screening for både bryst og livmoderhalskræft. Den tredje og nyeste kræftplan (2010) tilføjer et landsdækkende screeningsprogram for tyk- og endetarmskræft (KRC) med tilbud om screening hvert andet år til alle i alderen 50-74 år. Dette nye screeningsprogram skal iværksættes i 2014 [15][16, s. 1,7].

Et af de væsentligste problemer ved screeningsprogrammer er deltagelsesprocenten, da denne hænger tæt sammen med hvor mange tilfælde, der kan identificeres. Dette problem forstærkes af omkostningerne ved et screeningsprogram. Alene det første år forventes det nye screeningsprogram for KRC i Danmark at koste ca. kr. 200 millioner [17].
I to pilotstudier foretaget i Danmark i henholdsvis Vejle og København deltog kun 48% af de inviterede borgere, hvilket var en del under målet på 60%. På baggrund af studierne i Vejle og København anbefales det, at screeningsprogrammet kun indføres, hvis deltagelsesprocenten kan øges til omkring 60% [18]. Dette mål er sat ud fra international litteratur, som anbefaler en deltagelsesprocent på 60% eller mere [18, 19, 20]. Deltagelses procenten er med til at sætte fokus på identificeringen af alternative metoder til screening og detektering. Specielt er der fokus på biomarkører, der kan detektere KRC[3]. Flere review studier mener, at biomarkører er fremtiden for tidlig detektering af KRC, men de samme studier konkluderer samtidige, at der på nuværende tidspunkt ikke er fundet anvendelige biomarkører for detektering af KRC [3, 21, 22].

Dette giver anledning til at introducere en helt anden tankegang til at identificere biomarkører for tidlig detektering af KRC. I sundhedssektoren findes et stort antal informationssystemer [23]. Alle disse informationssystemer indeholder store mængder data, som primært er produceret i forbindelse med behandling af patienter. Det antages, at alle disse data har potentiale til at støtte sekundære formål, men dette arbejde har vist sig at være forbundet med nogle problemstillinger, som er svære at overkomme[24].
Anno 2013 har forfatterne, i et tidligere projekt, implementeret laboratoriedata fra LABKAII i Region Nordjylland (RN), i Regionens datavarehus, som i forvejen indeholdte data fra det patient administrative system (PAS) [1]. Dette gør det muligt at adressere problemstillinger som integrering af data og genbrug af data i en sekundær kontekst, fordi data fra to kildesystemer (PAS og LABKAII) er implementeret i samme datavarehus. Specielt er integreringen af diagnosekoder fra PAS og biokemiske analyser fra LABKAII interessant, fordi det gør det muligt at forbinde biokemiske analyser fra LABKAII med diagnosen, kolorektale kræft, registreret i PAS. Derigennem kan der opnåes et datasæt, som kan undersøges for potentielle kemiske biomarkører. Før datavarehuset, har denne type integrering af data, ikke været en mulighed i Region Nordjylland.
Vi foreslår derfor at genbruge data til detektering af KRC, ved at anvende Region Nordjyllands datavarehus til at lede efter sammenhænge i data. Genbrug af data og datavarehuset i sig selv er ikke et alternativ til screening, fordi genbrug af data kræver at borgerne har været i kontakt med sundhedssektoren. Men det kan måske bruges som et omkostningseffektivt alternativ til at analysere data fra de patienter, der har været i kontakt med sundhedsvæsnet og derigennem finde sammenhænge i data, der har potentiale til at kunne detektere KRC. Dette leder frem til følgende fire problemstillinger.

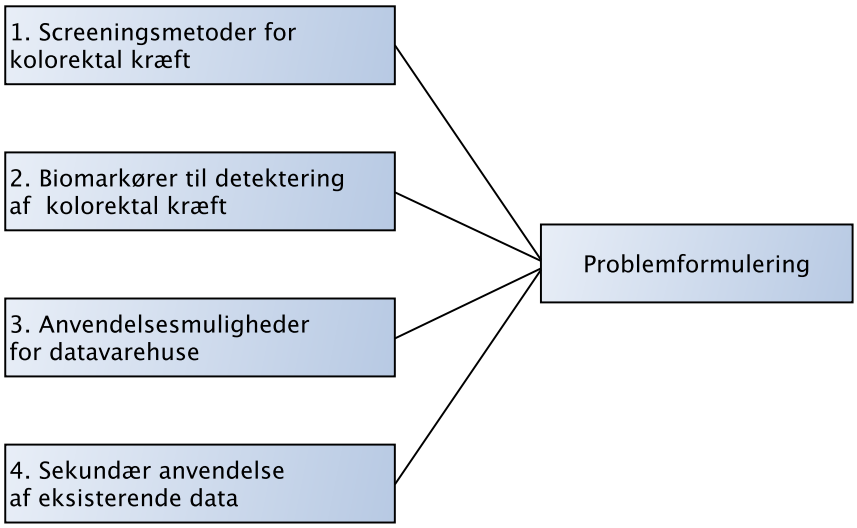
1. Ovenstående giver anledning til en nærmere undersøgelse af ulemper og fordele ved de nuværende screeningsmetoder, for at kunne sammenligne metoderne med resultaterne fra datavarehuset. Der er stor interesse i identificering af biomarkører, der kan være alternativer til de nuværende screeningsmetoder.
2. Hvis datavarehuset skal bruges til at identificere biomarkører, fordrer det en undersøgelse af hvorfor biomarkører har stor interesse i forbindelse med tidlig detektering af KRC.
3. Region Nordjyllands datavarehus har endnu ikke været anvendt til kliniske formål, som eksempelvis identificering af biomarkører til detektering af KRC. Dette giver anledning til at undersøge, hvilke anvendelsesmuligheder et datavarehus har med både kliniske data og patient administrative data.
4. Det at anvende et datavarehus er at anvende data til sekundære formål. Litteraturen introducerer en række problemstillinger i forbindelse med sekundær brug af data. Datavarehuset kan bruges til at adressere nogle af disse problemstillinger, og det er derfor nødvendigt at undersøge hvordan og derigennem hvilke forbehold, der skal tages ift. sekundær anvendelse af data.

De fire ovenstående problemstillinger leder frem til 4 spørgsmål, som igennem problemanalysen vil blive besvaret, inden den endelige problemformulering stilles. De fire spørgsmål er:

1. Hvilke fordele og ulemper er der ved nuværende screeningsmetoder?
2. Hvorfor er biomarkører i blodet et alternativ til nuværende screeningsmetoder?
3. Hvilke anvendelsesmuligheder giver et datavarehus?
4. Hvordan adresseres problemstillinger i forbindelse med sekundær brug af data gennem Region Nordjyllands datavarehus?

Problemanalyse 2

De fire indledende problemstillinger, der klarlægges i indledningen, omformes i problemanalysen til fire afsnit. Disse fire afsnit ligger til grund for og leder frem til den endelige problemformulering. Strukturen af problemanalysen kan ses på figur 2.1.
Det første afsnit er en analyse af ulemper og fordele ved en række screeningsmetoder. Der tages udgangspunkt i de screeningmetoder, der er blevet overvejet til det kommende screeningsprogram for kolorektal kræft (KRC) i Danmark.
Det andet afsnit undersøger, hvorfor biomarkører er interessante som alternativ til nuværende screeningsmetoder. Der tages udgangspunkt i tre review, og der gives 2 eksempler på biomarkører for KRC.
Det tredje afsnit præsenterer hvilke anvendelsesmuligheder, der er for et datavarehus i forbindelse med detektering af KRC. Der tages udgangspunkt i 3 internationale eksempler.
Det sidste afsnit behandler problemstillinger ift. sekundær anvendelse af data og hvordan disse problemstillinger adresseres i datavarehuset.



Figur 2.1: *Problemanalysens struktur*

2.1 Screeningsmetoder for Kolorektal Kræft

I indledningen blev følgende spørgsmål formuleret, og som vil blive besvaret i dette afsnit:

- Hvilke fordele og ulemper er der ved nuværende screeningsmetoder?

I 2012 er der i Danmark udarbejdet en MTV med anbefalinger til screeningmetoder som forberedende arbejde til det kommende screeningsprogram, der indføres fra 2014 [25]. I denne MTV anbefales det, at anvende screeningsmetoden iFOBT i det kommende screeningsprogram, fordi der er blevet fremlagt evidens for, at screening med iFOBT kan nedsætte dødeligheden af kolorektal kræft (KRC). Det anbefales også, at en positiv iFOBT følges op med en koloskopi [17]. I to gennemførlighedsstudier i hhv. Vejle og København er screeningsprogrammet blevet afprøvet. For begge studier var den samlede deltagelsesprocent under 50%. Studierne fandt frem til at den lave deltagelsesprocent skyldtes, at deltagerne syntes FOBT testen var besværlig at udføre selv, mens andre var nervøse for at få konstateret kræft [26]. Den lave deltagelsesprocent gør det interessant at undersøge andre fordele og ulemper ved de forskellige screeningsmetoder for KRC, fordi det kan påvirke screeningsprogrammerne.

Det følgende kapitel vil derfor analysere og sammenholde et udvalg af de hyppigst anvendte screeningsmetoder til KRC med henblik på at belyse fordele og ulemper. De hyppigste screeningsmetoder for KRC, som også er nævnt i de danske anbefalinger, omfatter [3, 8, 27]:

- Fækal Okkult Blodtest (FOBT)
- Fækal DNA Test
- Fleksibel Sigmoideoskopi
- Koloskopi
- Virtuel Koloskopi / CT Koloskopi

Fækal Okkult Blodtest (FOBT)

FOBT er en af de anbefalede screeningsmetoder for KRC, og i Europa er det den mest anvendte. Dette skyldes blandt andet, at den er blevet valideret i *randomized control trials*, den skulle være nem at udføre og den er billig. Der findes to forskellige principper for FOBT screening henholdsvis kaldet gFOBT, der er guaiac-baseret og iFOBT, som er en immunkemisk test. Fælles for begge principper er, at de begge bruges til at påvise tilstedeværelsen af ikke synligt (okkult) blod i fæces [9, 8]. Fælles for FOBT analyserne er, at det i screeningsprogrammer anbefales at personer, der er 50 år eller ældre og er i risikogruppen for KRC, tilbydes screening med FOBT hvert andet år [28, 29, 16].

gFOBT metoden blev allerede beskrevet i 1864 men først udviklet i 1967-1970, hvor den blev markedsført, hvilket gør gFOBT til en af de ældste af screeningsmetoderne for KRC. Metoden er baseret på reaktionen mellem guaiac og peroxidase. Peroxidase findes i hæmoglobin, og det giver en blå farve når det reagere med guaiac. Derfor vil testen blive blå ved tilstedeværelsen af blod i afføringsprøven. En af fordelene ved gFOBT er, at det er en simpel test, som kan udføres ved den praktiserende læge på mindre end 2 minutter. Men på trods, af at gFOBT er en simpel test, har den en række ulemper sammenlignet med iFOBT. For det første kræver testen 6 forskellige fæcesprøver fra tre forskellige afføringer. Derudover er gFOBT usikker over for indtag

af forskellige fødevarer, hvorfor det er nødvendigt med diætrestriktioner 72 timer inden der opsamles fæcesprøver. Diæten afgrænses inden prøveopsamling til ikke at indeholde, rødt kød (lam, oksekød, lever), peberrod, radiser, ukogte eller let kogte roer, broccoli, blomkål, melon, spinat, banan, tomat, pastinak eller alkohol, da disse fødevarer indeholder peroxidase og kan dermed give falske positive resultater. Fødevarer med et rigt indhold af C-vitamin kan give falske negative resultater, når gFOBT anvendes.
Foruden fødevarerestriktionerne kan medicin i form af cortico steroider, kemo, aspirin og andre non-steroide antiinflammatoriske stoffer resulterer i gastrointestinalt blodtab, hvilket ligeledes kan give falske positive resultater [8]. Da der er forskellige producenter af gFOBT, afhænger specificitet og sensitivitet for KRC af de enkelte præparater.
Specificiteten for KRC med en gFOBT er fundet til at ligge i intervallet 96-99%, mens sensitiviteten varierer fra 9% til 81% [8, 30, 21]. Sensitivitet siger noget om hvor god metoden er til at identificere alle de syge patienter. Dvs. ved 100% sensitivitet findes alle syge patienter og ingen syge patienter klassificeres som raske. Sensitiviteten siger ikke noget om hvor mange patienter, der klassificeres som syge selv om de er raske, det gør specificiteten. I et screeningsperspektiv er det problematisk at overse syge patienter, dvs. fortælle syge patienter at de faktisk er raske [31]. Derfor er sensitiviteten meget vigtig, fordi man er interesseret i at finde alle syge patienter, selv om man også finder nogle patienter, som er raske.

Den anden variant af FOBT er iFOBT, og den er mere præcis end gFOBT, da den reagerer specifikt på intakte globinmolekyler fra humant hæmoglobin. Dette giver færre falske positive resultater, sammenlignet med gFOBT, da mad ikke giver udslag. Reduktionen i falske positive, mindsker udgifterne til unødvendige opfølgninger. iFOBT kræver kun to forskellige afførringsprøver, men kan udføres med en enkelt afføringsprøve. Men på grund af at tumorer kan have afbrudt blødning, opnåes der bedre resultater med to prøver [21, 8]. En fordel men også en ulempe ved iFOBT er at prøven skal analyseres af et laboratorium, og dermed ikke kan aflæses af den enkelte læge, som er tilfældet med gFOBT. Yderligere har iFOBT heller ikke nogen restriktioner i forhold til medicin og diæt, hvilket gør testen nemmere at udføre for patienten [8, 9, 21]. Specificiteten for KRC med iFOBT er fundet til at ligge i intervallet 86-98% og sensitiviteten ligger i intervallet 62-100% [8, 30].

Fækal DNA Test

Foruden iFOBT og gFOBT findes der også en fækal DNA test til at detektere KRC. DNA testen består reelt af 23 forskellige tests, som detekterer flere forskellige cellemarkører fra præmalignante adenomer og tumorer. DNA testen kræver en enkelt afføringsprøve på minimum 30g, som sendes til et laboratorium, der udfører de procedurer, der skal til for at udtrække humant DNA fra prøven, og derefter analyserer det for kræftmarkører. Det tager normalt 2-3 uger fra DNA prøven er sendt, til der foreligger et resultat. DNA testen for KRC er desuden en dyr test sammenlignet med FOBT. DNA test for KRC har en fordel over FOBT, da forstadier til kræft kan detekteres. Testen har ligesom iFOBT heller ingen diæt eller medicin restriktioner, hvilket er en fordel for patienten [8]. Specificitet for KRC med DNA testen, ligger i intervallet 93-100%, mens sensitiviteten er i intervallet 71-91%. Desuden har DNA testen en sensitivitet for adenomer i intervallet 55-82% [8, 32]. DNA testen kan have falske negative resultater, da kræfttumorer er genetisk heterogene, og fordi ingen DNA mutation er konsistent i alle typer af tumorer. DNA testen kan have falske positive resultater, pga. kræft i bugspytkirtlen, maven, spiserøret eller i lungerne [8].

Fleksibel Sigmoideoskopi

Fleksibel sigmoideoskopi er en endoskopisk undersøgelse af endetarmen og den nedre tyktarm. Proceduren udføres enten med et sigmoideoskop, som er 60 cm langt. Inden proceduren kan udføres på patienten, er det nødvendigt at rense nedre tyk og endetarm, hvilket gøres med et lavement, der administreres 30-60 minutter før undersøgelsen. Sigmoideoskopiundersøgelsen er begrænset af sigmoideoskopets længde, hvor godt tarmen er blevet renset og patientens tolerance, da undersøgelsen normalt udføres uden bedøvelse. Fleksibel sigmoideoskopi har en høj sensitivitet og specificitet for både adenomer og KRC, i det område hvor sigmoideoskopet kan undersøge. Ved sigmoideoskopi kan 80% af KRC opdages, da sigmoideoskopet kan undersøge hele endetarmen og hele den venstre side af tyktarmen.[9, 11] Udførelse af sigmoideoskopi kræver højtuddannet kompetent personale og gode kvalitetsikringsprocedurer for at opnå gode resultater. Undersøgelsen har desuden en lille risici for større komplikationer, da det er en invasiv undersøgelse. Komplikationerne omfatter blandt andet blødning pga. perforation af tarmen. Denne risici stiger ved opfølgende koloskopi, hvor der fjernes polypper. Sigmoideoskopi er til sammenligning med FOBT, betydeligt bedre til at detekterer højrisiko adenomer og KRC. Screening med fleksibel sigmoideoskopi anbefales hvert 5. år til personer, der er 50 år eller ældre [9, 11, 32, 28].

Koloskopi

Koloskopi anses for at være den gyldne standard til undersøgelse af tyk og endetarm. Koloskopi er ligesom sigmoideoskopi en endoskopisk undersøgelse, der udføres med et koloskop. Et koloskop er 130-160 cm langt. Ligesom med sigmoideoskopi kræver koloskopi også at patientens tyk og endetarm renses inden undersøgelsen udføres, samt at patienten bedøves.
Udførelse af koloskopi kræver højtuddannet kompetent personale samt gode kvalitetsikringsprocedurer. Dette, kombineret med at patientens forberedelse til undersøgelsen og bedøvelsen af patienten, er med til at øge omkostningerne ved proceduren. Koloskopi giver desuden ligesom sigemoideoskopi også risici for blødning og perforation af tarmen i forbindelse med undersøgelsen. Til gengæld har koloskopiundersøgelsen en højere sensitivitet og specificitet for både adenomer og KRC end de ovenstående undersøgelser, da det er muligt at undersøge hele patientens tyk- og endetarm. Screening med koloskopi anbefales hvert 10. år til personer, der er 50 år eller ældre[9, 11, 32, 28, 29]. I ca. 1:1000 tilfælde medfører koloskopi komplikationer for patienten[33]

Virtuel Koloskopi

En af de nyere muligheder for at screene for KRC, er virtuel koloskopi. Virtuel koloskopi er en Computed Tomography (CT) scanning af tyk og endetarm. Dette er en ny metode, som udviser gode muligheder for screening især for store adenomer, hvor de indledende studier viser en sensitivitet og specificitet på niveau med koloskopi. Fordelen ved CT kolografi er, at det er en hurtig og non-invasiv metode. Men den kræver stadig rensning af tarmen ligesom de to ovenstående metoder. Desuden kræver undersøgelsen CT scannere og software, der er dyrt. Ved CT scanning udsættes patienten for stråling, som kan føre til senere komplikationer. Derudover mangler Virtuel Koloskopi som screeningsmetode evidens[32, 11].

Metode	Fordele	Ulemper
gFOBT	+ Non-invasiv + Nem at udføre + Manuel aflæsning + Lav stk. pris + Høj specificitet	- Diæt Restriktioner - Medicin Restriktioner - Manuel aflæsning - 3 afføringsprøver - varierende sensitivitet
iFOBT	+ Non-invasiv + Maskinaflæsning + Ingen Diæt Restriktioner + Ingen Medicin Restriktioner + Nem at udføre + Lav stk. pris	- Skal sendes til laboratorium - Maskinaflæsning - Varierende sensitivitet
Fæces DNA test	+ Non-invasiv + Ingen Diæt Restriktioner + Ingen Medicin Restriktioner + Nem at udføre + Lange screeningsintervaller	- Skal sendes til laboratorium - Svartid 2-3 uger - Dyr at udføre
Sigmoideoskopi	+ Høj specificitet + Høj sensitivitet + 80% af KRC opdages	- Invasiv - Blødning og perforering af tarm - Dyr at udføre - Kræver højtuddannet personale - Ubehagelig for patienten
Koloskopi	+ Høj specificitet + Høj sensitivitet + >80% af KRC opdages + Langt mellem screeninger	- Invasiv - Blødning og perforering af tarm - Dyr at udføre - Kræver højtuddannet personale - Ubehagelig for patienten - Tyk og endetarmsrensning - Bedøvelse
Virtuel Koloskopi (CT)	+ Non-invasiv + Hurtig udførelse	- Meget dyr i anskaffelse - Manglende evidens - Tyk og endetarmsrensning - Dyrt udstyr

Tabel 2.1: *Opsummering på fordele og ulemper på de forskellige screeningsmetoder for KRC.*

2.1.1 Delkonklusion

Tabel 2.1 viser en opsummering af fordele og ulemper ved hver af de forskellige screeningsmetoder. Koloskopi, der anses for at være den gyldne standard indenfor KRC screening, anbefales til personer, der er 50 år eller ældre, og som har høj risiko for at udvikle KRC. Men koloskopi er en omkostningsfuld og invasiv procedure, som i ca. 1:1000 tilfælde også medfører komplikationer [33]. Det er forbundet med ubehag for patienten at få foretaget både koloskopi og sigmoideoskopi, hvorfor der i screeningsprogrammer, hvor disse metoder hovedsageligt anvendes, er en dårlig deltagelsesprocent [3, 29]. Sigmoideoskopi ligner koloskopi og har derfor de samme fordele og ulemper som koloskopi, men sigmoideoskopi giver ikke helt den samme høje sensitivitet og specificitet som koloskopi, da sigmoideoskopi kun kan bruges til at undersøge den venstre side af tyktarmen, hvorved kun 60-70% af adenomer og kræft kan detekteres.

DNA testen er omkostningsfuld, og det kan tage op til 3 uger at få svar på testen.
iFOBT og gFOBT har begge meget varierende sensitivitet, hvilket kan skyldes, at tumorer kan bløde sporadisk. Derfor øges sensitiviteten også ved at teste flere fæcesprøver fra en patient ad gangen. Lav sensitivitet kan være problematisk i et screeningsperspektiv, fordi deltagere kan risikere at blive klassificeret som raske, selvom de er syge.
Virtuel Koloskopi udsætter patienterne for stråling og metoden mangler afgørende evidens.

2.2 Biomarkører til detektering af kolorektal kræft

Forrige afsnit om screening afslørede, at de nuværende metoder til screening enten har varierende sensitivitet og specificitet, er ressourcekrævende eller invasive, som gør at deltagelsesprocenten påvirkes negativt. I litteraturen er der derfor stor fokus på at finde alternativer. Specielt biomarkører i blodet til detektering af kolorektal kræft (KRC) får stor interesse, hvilket i indledningen ledte frem til spørgsmålet:

- Hvorfor er biomarkører i blodet et alternativ til nuværende screeningsmetoder?

Dette afsnit giver en definition af biomarkører samt eksempler på biomarkører, som undersøges i forbindelse med KRC.

Ofte defineres biomarkører som en substans i forbindelse med kræft, der kan bruges til at indikere en biologisk tilstand. Biomarkører bruges til at evaluere normale biologiske processer, patogenese og effekten af medikamenter, samt at detektere sygdomme. Biomarkører bruges f.eks. allerede til at finde den optimale behandling ift. forskellige kræfttyper.
Biomarkører kan være både kemiske, fysiske og biologiske, hvor biomarkører identificeret i en blodprøve ofte er kemiske[3, 34]. Der vil i dette projekt være fokus på kemiske biomarkører, fordi litteraturen fokusere på biomarkører i blodet[3, 34].
Der er to typer af biomarkører, som det er vigtigt at skelne imellem. Der er biomarkører, som kan indikere risikoen for at udvikle en sygdom i fremtiden givet, at patienten er udsat for en eller flere risikofaktorer. Og der er biomarkører, som indikerer at patienten har en sygdom eller er ved at udvikle sygdommen, hvis en faktor måles i kroppen. Den sidste type biomarkør bruges bl.a. til screening og til at udvikle målrettede behandlinger. Det er denne type biomarkører, som der vil være fokus på i dette afsnit, netop fordi den bruges i forbindelse med screening og detektering.

Langt de fleste studier, der identificerer biomarkører, undersøger proteiner [3, 34], fordi proteiner påvirkes af muterede gener[35]. [34] identificerer 70 forskellige biomarkører fra blodprøver, i 93 forskellige studier, mens [3] identificerer 46 forskellige studier. Begge review konkluderer dog, at der indtil nu har været begrænset succes med at identificere biomarkører [3, 34]. Mange studier viser at biomarkøren, som undersøges, er knyttet til udviklingen af KRC, men studierne har svært ved at demonstrere den ønskede diskriminative karakter i forhold til KRC. Dog er interessen for biomarkører i blodet stadig stor, fordi den slags biomarkører kan give hurtig og minimalt invasiv detektering af KRC [3, 22].
[3] identificerer, at de fleste studier undersøger en eller meget få biomarkører ad gangen (univariate analyser), og foreslår at det kan være grunden til den begrænsede succes. Derfor foreslås det, at relevante biomarkører kombineres til et panel af biomarkører, da carcinogenese af KRC er så kompleks, at evaluering af én biomarkør ad gangen ikke kan

beskrive kompleksiteten.
Dette underbygger beskrivelsen af kræft, som den store efterligner[10, s. 605-607] og er med til at gøre det sandsynligt, at univariate dataanalyser er for simple til at beskrive kræft. For at understrege denne pointe kan en parallel drages til klinisk praksis, hvor en diagnose sjældent stilles på baggrund af en enkelt faktor. Under diagnoseprocessen indsamles flere og flere informationer indtil det er muligt at af- eller bekræfte en diagnose [36].
Følgende underafsnit giver to eksempler på proteiner, hvis evne til at indikere KRC er blevet evalueret.

C-Reaktive Protein (CRP)

I review'et af [3] undersøges CRP i 13 studier ud af 46. Et af studierne tester CRP-niveauet i plasma. Hypotesen er, at CRP-niveauet er højere for patienter med KRC end normal værdien [37]. Studiet finder et forhøjet niveau hos patienter med KRC, hvilket også er forventeligt, da det antages at inflammation spiller en rolle ift. udviklingen af KRC. Studiet foreslår derfor at CRP kan bruges til at identificere patienter med KRC, men også at mange andre sygdomme påvirker CRP.

Ferritin

Et studie i Mexico City undersøger ferritin ift. FOBT-testen, som er beskrevet i forrige afsnit 2.1. Ferritin er et protein i plasma, som er et depot for jern i kroppen. Ferritin undersøges, fordi det antages at tumorer i tarmen bløder, og derfor kan patienter med KRC opnå en anæmisk tilstand, hvor ferritin-niveauet falder.
Alle 57 patienter i studiet havde en positiv FOBT-test. Ud af 57 havde 21 KRC, og 7 havde adenomer. Studiet fandt ingen forskel i ferritin niveauet mellem patienter med KRC, adenomer og patienter med en negativ koloskopi. Men studiet fandt samtidige at albumin var signifikant lavere hos patienter med KRC end hos patienter uden KRC, men ingen konklusion drages fra dette, da dette ikke var artiklens fokus [38].

2.2.1 Delkonklusion

På grund af ulemper ved de nuværende screeningsmetoder for KRC, fokuseres på at identificere biomarkører i blodet, fordi de er hurtige at udføre og minimalt invasive. To review identificerer tilsammen over 100 biomarkører, men trods denne store mængde, er der endnu ikke fundet evidens for biomarkører, som kan diskriminere KRC. Trods det er flere biomarkører kædet sammen med udviklingen af KRC, hvorfor biomarkører har stor interesse. Derudover foreslår et af review studierne, at kombinationen af to eller flere biomarkører til et panel af biomarkører, kan øge detekteringspotentialet.

2.3 Anvendelsesmuligheder for datavarehuse

I indledningen blev følgende spørgsmål omkring anvendelse af datavarehuse formuleret.

- Hvilke anvendelsesmuligheder giver et datavarehus?

Dette vil blive besvaret i det kommende afsnit.

Region Nordjylland's (RN) datavarehus hedder FaktaCenter, og det indeholder patientadministrative data og laboratoriedata. Se appendiks B på side 85 for mere information om FaktaCenteret og den underliggende model. Datavarehuset er endnu ikke blevet anvendt i forbindelse med klinisk forskning eller andre kliniske formål.
Dette afsnit præsenterer først, hvilke anvendelsesmuligheder litteraturen beskriver for datavarehuse. Derefter gives nogle eksempler på datavarehus-projekter og de formål, som datavarehusene anvendes til.

Et datavarehus er en homogen datakilde, som indeholder data fra en eller flere kilder, med det formål at data kan bruges til sekundære formål. Helt grundlæggende er målet med et datavarehus, at stille data til rådighed for beslutningstagere. Men det at stille data til rådighed er ikke nok. Data skal også omdannes til information, som kan understøtte et formål, og dette kan ske ved f.eks. data mining og/eller beslutningsstøtte[39]. I litteraturen beskrives tre overordnede formål, som informationen fra et datavarehus kan understøtte [40]:

- Kliniske formål.
- Administrative formål.
- Kvalitetssikrings formål.

Kliniske formål omfatter blandt andet: Rekruttering af patienter til kliniske forsøg, finde tendenser relateret til befolkningens sundhed, brugen af, effekten af og prisen på medikamenter, samt identificering af faktorer, der leder til sygdomme, infektionsmonitorering og hurtigere diagnosticering [41].
Administrative formål omfatter blandt andet: Planlægning af kapacitet, planlægning af produktindkøb, afregning og patient bevægelse igennem sygehuset [41].
Kvalitetssikrende formål omfatter bl.a. beslutningsstøtte til at undgå fejl, samt at sikre kvalitet i behandlingen og dokumentation[41].

Da identificering af biomarkører til detektering af kolorektal kræft (KRC) er en kliniske problemstilling, vil resten af dette afsnit have et kliniske omdrejningspunkt. FaktaCenteret i RN indeholder patientadministrative data, som normalt vil falde under kategorien administrative data. FaktaCenteret indeholder også laboratoriedata, som falder under kategorien kliniske data. Følgende tre underafsnit vil give eksempler på de formål, som data fra forskellige datakilder tilsammen kan understøtte.

Care Data Warehouse

I Sverige i Östergötalands kommune findes et datavarehus, som kaldes *Care Data Warehouse* (CDW). Det indeholder patientadministrative data [42], så som information om en kontakt til sundhedssektoren inklusiv en diagnosekode. Denne database refereres flere gange i litteraturen, som et værktøj til at identificere patienter, der skal inkluderes i studier [43] eller til at studere prævalensen og omkostningerne forbundet med sygdomme [44]. CDW bruges altså til at levere beslutningsstøttende information, hvad enten det er hvilke patienter, der skal inkluderes i et studie eller sygdommes indvirkning på samfundet. CDW indeholder kun patientadministrative data.

2.3.1 Chicago Antimicrobacterial Resistance Project

Udenfor Skandinavien ses andre datavarehus projekter, som anvender både administrative og kliniske data. Et eksempel fra Chicago benytter data fra både radiologiske, medicinske, patientadministrative og laboratorie-informationssystemer, samt elektroniske patientjournaler til udviklingen af et informationssystem til infektionskontrol. Projektet kaldes *Chicago Antimicrobacterial Resistance Project* (CARP) [45]. Formålet var at monitorere hospitalsinfektioner for dermed bedre, at kunne kontrollere og behandle infektionsudbrud. Integrationen af kliniske og administrative data er blevet brugt til både klinisk beslutningsstøtte og administrativ effektivisering i forbindelse med hospitalsinfektioner. Resultatet var både en højnet behandlingskvalitet og en effektivisering af ressourceforbruget ift. behandling og kontrol af infektionsudbrud.
Forfatterne kalder dette informationssystem for et datavarehus. Informationssystemet blev designet som et infektionsinformationssystem, med en relationel datamodel. Relationelle modeller er ofte svære at udvide, når de først er etableret pga. af den 3. normalform, som traditionel relationel modellering forsøger at opnå[46]. Systemet i dette eksempel er ikke komplet, da det kun indeholder infektionsrelevante data. Denne karakteristika gør at dette informationssystem mere ligner en avanceret klinisk database end et datavarehus. Datavarehuse laves ikke som silosystemer til et enkelt formål. I princippet bør den underliggende datavarehusmodel omfavne enten alle data i organisationen eller muligheden for, at det kan udvides med alle data[47]. Derudover bør datavarehuse ikke udvikles til et specifikt formål. For at sætte det på en spids svarer det til at udvikle et nyt datavarehus, hver gang man vil lave en ny forespørgsel. Datavarehusmodellen bør laves, så den kan omfavne alle relevante data fra kildesystemet. Data i den mere generelle model kan derefter bruges til specifikke formål som f.eks. infektionskontrol.
Pointen er, at "datavarehuse/klinisk databaser lavet til et specifik formål kan bruges til meget begrænsede formål, mens et datavarehus med en mere generel datamodel kan bruges til at støtte mange formål[46].

2.3.2 Intermountain Healthcare's enterprise datavarehus

Et godt eksempel på et datavarehus, som har en mere generel datamodel, er enterprise datavarehuset bygget af organisationen *Intermountain Healthcare* i Utah og Idaho. Den underliggende model indeholder de fleste data fra Intermountain Healthcare organisationen, og modellen er ikke bygget til et specifikt formål, som det tidligere eksempel. *Intermountain Healthcare's* datavarehus er blevet bygget til mere generelle formål, som bl.a. omfatter dokumentation af *best practice*, økonomi og kvalitetssikring. Dette datavarehus har været brugt i flere meget forskellige studier. Bl.a. er datavarehusets effekt, som alarmsystem overfor

multiresistente bakterier, blevet studeret. Derudover er datavarehuset også blevet brugt til at identificerer risikofaktorer for dyb venetrombose [41].
Intermountain Healthcare har stor fokus på at udnytte datavarehuset i forbindelse med beslutningsstøtte. Det er bl.a. andet i forbindelse med prædiktering af venetrombose. Dette eksempel udnytter datavarehuset ved at implementere alarmsystemer. Eksemplet udnytter også data mining til at identificere risikofaktorer.

Datavarehus	**Datatyper**	**Anvendelse**
Care Data Warehouse Sverige	-Patientadministrative data	-Identificere forsøgspersoner -Prævalens studier
CARP	-Patientadministrative data -Medicineringsdata -laboratoriedata -Radiologiske data -Infektionsundersøgelser	-Forbedring af infektionsrelaterede problemstillinger
Intermountain Healthcare EDW	-Patientadministrative data -National mortalitetsdatabase -Laboratoriedata -Mikrobiologidata -Medicineringsdata -EPJ	-Identifiere risikogrupper for venetrombose -Forbedring af infektionsrelaterede problemstillinger -Identificere risikofaktorer -Værktøj til beslutningstøtte -Værktøj til data mining
FaktaCenter	-Patientadministrative data -Laboratoriedata	-Endnu ikke anvendt til kliniske formål

Tabel 2.2: *Opsummering af hvilke datatyper de fire datavarehuse indeholder, og hvad de er blevet anvendt til.*

2.3.3 Delkonklusion

Først og fremmest er det ikke nok at have et datavarehus med data fra flere kilder. Denne samling af data skal omsættes til information. Flere studier har foreslået, at datavarehuse kan bruges til beslutningsstøtte og data mining ift. både administrative formål, kliniske formål og kvalitetssikring. Flere eksempler i litteraturen understøtter påstanden, ved bl.a. at demonstrere effekten af beslutningsstøtte og data mining i forbindelse med eksempelvis kontrollering af infektionsudbrud og identificering af risikofaktorer for dyb venetrombose.

RN's datavarehus indeholder nogle af de samme datatyper, som de øvrige datavarehuse i tabel 2.2. Der er endnu ikke anvendt beslutningsstøtte eller data mining i RN's datavarehus i forbindelse med kliniske problemstillinger, for at omdanne data til information om detektering af KRC, men flere andre datavarehuse demonstrerer gode anvendelsesmuligheder ved at kombinere flere forskellige datakilder. Det understøtter muligheden for at identificere biomarkører for KRC i RN's datavarehus.

2.4 Sekundær anvendelse af eksisterende data

Anvendelse af datavarehuset, der indeholder data opsamlet igennem de primære processer i sundhedssektoren, til detekteringen af kolorektal kræft (KRC) indebærer sekundær brug af eksisterende data. Med andre ord siges det, at data kan genbruges til andre formål end det/de formål, som data oprindeligt blev opsamlet til. Flere studier har demonstreret, at sekundær brug kan skabe stor værdi [45, 41, 43]. Der er dog nogle problemstillinger forbundet med sekundær brug af data, som i sundhedssektoren har vist sig at være svære at overkomme. Begge problemstillinger herunder vil først blive præsenteret i dette afsnit, før spørgsmål 4 fra indledningen besvares. Problemstillingerne, identificeret i litteraturen, ift. sekundær brug er:

1. Data er kontekst afhængige [24].
2. Mange datakilder er enkeltstående informationssystemer uden integrering[40].

1. Konceptet i at genbruge data til forskning indebærer, at individuelle data omfortolkes til generel information[48]. Denne proces besværliggøres, fordi individuelle data er tæt knyttet til den primære kontekst, hvori de er produceret. Denne kontekst skal håndteres, hvis individuelle data skal omdannes til generel information. F.eks. kan laboratoriedata ikke fortolkes uden viden om laboratoriets normalværdier.
Udredningen af data bliver mere og mere ressourcekrævende, desto flere sekundære formål data skal understøtte [24]. F.eks. er der forskel på manglende data og irrelevante data. Ofte opsamles kun relevante data for en given klinisk problemstilling eller patient. Relevans er ofte subjektivt. I begge tilfælde er data ikke registrerede, men det kan være svært at gennemskue hvilket tilfælde, der er gældende for et givent patientforløb uden en kontekst. I klinisk kontekst fortæller den samlede mængde af data en historie om patientens tilstand, som ikke er tydelig, hvis dataelementerne ses enkeltvis. Når mere data tilføjes, genfortolkes data til en ny historie. Gamle data kan få ny mening når nye data tilføjes, hvilket især kan gøre klassificering af data problematisk. Det kan dog lade sig gøre at udrede data fra konteksten, men det kræver mange ressourcer [24].

2. Ofte er informationssystemerne ikke integrerede, og databaseimplementeringerne følger ikke fælles standarder [23]. Dvs. at systemerne har forskellige databasemodeller med forskellig forespørgselssyntaks og systemerne understøtter forskellige kliniske terminologier, hvilket gør det besværligt at samle og/eller aggregere data på tværs af systemerne.
Forskellighederne i informationssystemerne resulterer i et stort antal heterogene databaser, som ikke kan udveksle data. Det er i sig selv ikke et problem ift. sekundær brug af data, hvis et enkelt informationssystem indeholdte alle de data, som var nødvendige for sekundære formål. Ofte er dette ikke tilfældet, fordi informationssystemerne er implementeret til et bestemt formål [23]. Det betyder, at de indeholder forskellige typer af data afhængigt af det primære formål. F.eks. indeholder PAS administrative informationer, såsom diagnoser, patientkontakter, procedurer, sengedage med videre. Der er med andre ord meget få kliniske informationer, fordi PAS har administrative formål. Det gør det mindre relevant at bruge data fra PAS til f.eks. klinisk forskning.
LABKAII, som er et laboratorieinformationssystem, indeholder prøvesvar på bl.a. blod- og urinprøver. LABKAII indeholder ingen administrative informationer om patienten udover det patient ID, som prøvesvarene tilhører. Dette er ikke en hindring ift. det primære formål, men det begrænser mulighederne for klinisk forskning, fordi prøvesvaret ikke er kædet sammen med

relevante administrative informationer så som aktions-, henvisnings og tillægsdiagnoser. Dette giver ideen til at integrere kliniske og administrative data.

Hvis datavarehuset skal anvendes til det sekundære formål at identificere biomarkører for KRC, er det vigtigt at vide, hvordan disse problemstillinger overkommes. Det leder frem til følgende spørgsmål, som besvares i dette afsnit:

- Hvordan adresseres problemstillinger i forbindelse med sekundær brug af data igennem Region Nordjyllands datavarehus?

Før data implementeres i et datavarehus eller undervejs i implementeringen udføres en stor del af arbejdet med at udrede data fra konteksten og klargøre data til integration.
[48] beskriver hvordan klassifikationssystemer kan bruges til at udrede data fra den primære kontekst (1), så data bliver sammenlignelige. Et Klassifikationssystem indeholder unikke klasser, som ikke overlapper hinanden. Hvilket vil sige at klasserne kan optælles i forbindelse med statistiske formål. Da klasserne er unikke, så indeholder klassifikationssystemer beskrivelser af, hvordan en klasse tildeles. Dvs. at der er information om hvad, der skal undersøges før en klasse tildeles. At en klasse skal tildeles unikt, gør klassifikationssystemer meget stærke ift. statistiske formål, fordi klasserne kan sammenlignes uden, at der opstår overlap i data. Det er et stort arbejde at klassificere data. Ofte skal der opsamles flere data for, at kunne tildele data en unik klasse, end hvad der er behov for til det primære formål.
Klassificering gør det muligt igennem statistiske dataanalyser, at omfortolke individuel information til generel information.

I Danmark gøres et stort stykke arbejde med at klassificere data. Patientadministrative systemer (PAS) på alle sygehuse i Danmark, rapporterer til bl.a. landspatientregisteret, som kræver at data er klassificeret. Data bliver klassificeret efter SKS-systemet, der bygger på ICD10 og NCSP [49]. Dvs. at patientadministrative data fra PAS i RN's datavarehus er klassificeret med SKS-koder. Ydermere er laboratoriedata i LABKAII klassificeret med NPU-terminologien [50]. Dette gør statistisk evaluering mulig, fordi dataopsamlingen guides af regler for, hvordan en klasse tildeles.

RN's datavarehus adresserer integreringsproblemstillingen (2), ved at implementere data fra de to datakilder i en fælles underliggende dimensionel model. Se appendiks B på side 85 for mere information om datavarehusets dimensionelle model. Det betyder, at de to kildesystemer deler tabeller, der hvor der er overlap i data. F.eks. eksisterer CPR-nummeret for patienten i begge kildesystemer[1].

[48] forklarer, hvordan data kan omdannes fra, at kunne bruges til individuel behandling af patienterne til forskning. Denne proces viser klassifikation af data, som den sidste opgave inden data kan omdannes til generel information. Den sidste opgave, der omdanner data til information, er statistisk dataanalyse. [48] illustrer, at blot fordi data er klassificeret og gjort klar til integration, betyder det ikke, at data automatisk bliver til information. Forbindes dette argument med udsagnet fra [24], om at data skal udredes mere og mere fra den oprindelige kontekst, desto længere væk data tages fra den oprindelige kontekst, så betyder det, at den sidste cyklus, hvor data omdannes fra komparative data til generel information, stadig

inkluderer integrations- og kontekst-afhængighedsproblemstillingerne. Dette er en vigtigt pointe ift. til den kommende dataanalyse, og disse problemstillinger skal overvejes grundigt i metoden. Dermed bliver datavarehuset ikke en direkte løsning til de to problemstillinger, men datavarehuset bliver et unikt værktøj, som kan bruges til at udrede data fra konteksten og integrere data til et datasæt, der kan bruges til sekundær anvendelse.

2.4.1 Delkonklusion

Der er identificeret to problemstillinger i litteraturen i forbindelse med sekundær anvendelse af data. 1. Data er kontekstafhængige, hvilket betyder, at data skal udredes fra den primære kontekst, som de er produceret i, før data kan genbruges. 2. Mange datakilder er ikke integrerede. Hvis det sekundære formål kræver data fra flere kilder, kræver dette integrering.

RN's datavarehus med patientadministrative data og laboratoriedata, adresserer begge problemstillinger. Konteksten adresseres gennem klassifikation af data, og integrationen adresseres ved hjælp af datavarehusets underliggende model.
Dermed er et stort arbejde allerede klaret i forbindelse med sekundær anvendelse af data, til identificering af kemisk biomarkører til detektering af KRC. Den sidste opgave i processen, hvor data omdannes til information, betyder at problemstilling 1) og 2) ikke er løst vha. datavarehuset, men at metoden til dataanalysen skal tage højde for både konteksten i og integreringen af data.

2.5 Problemformulering

Kolorektal kræft (KRC) er den tredje hyppigste kræftform og den næst hyppigste dødsårsag blandt kendte kræftformer. Tilmed er den 5-årige overlevelsesrate, efter diagnosen er stillet, i gennemsnit 50%. Men studier har vist, at tidlig diagnose og tidlig indsats kan øge den 5 årige overlevelsesrate til 90%. Disse uoverensstemmelser gør KRC til et emne, hvor der er behov for screening eller detektering, så tidlig indsats faciliteres.
I Danmark iværksættes et screeningsprogram i 2014, men indledende undersøgelser har vist, at deltagelsesprocenten er under 50%. Derudover har den internationale litteratur fundet, at de nuværende screeningsmetoder enten har meget varierende sensitivitet og specificitet eller er dyre, besværlige og ubehagelige for patienten. Ulemperne og den lave deltagelsesprocent betyder, at der ledes efter alternative metoder til de nuværende metoder. Kemiske biomarkører, specielt i blodet, har i litteraturen stor interesse, fordi de har potentiale for at give hurtig og minimal invasiv detektering af KRC. Der vil derfor være fokus på at identificere kemiske biomarkører i dette projekt. Der er endnu ikke fundet evidens for biomarkører til detektering af KRC. Trods det er interessen stadig stor, fordi flere af de undersøgte biomarkører er kædet sammen med udviklingen af KRC. De har dog ikke kunne detektere KRC med biomarkørerne. Derudover blev det foreslået, at flere biomarkører kunne kombineres til et panel, som bedre beskriver den komplekse carcinogenese af KRC.

Da laboratoriedata er blevet implementeret i Region Nordjylland's (RN) datavarehus med patientadministrative data, så giver det en datakilde med adgang til biokemiske analyser, der kan undersøges som kemiske biomarkører for KRC. RN's datavarehus giver adgang til store mængder data, dermed kan data mining bruges til at identificere kemiske biomarkører, ligesom der i andre datavarehus-studier er blevet anvendt data mining. De store mængder data gør også, at datavarehuset har potentiale til at understøtte dataanalyser, der kombinerer flere biokemiske analyser.
Datavarehuset skal ikke ses som et direkte alternativ til screening, fordi det kun er patienter, der har været i kontakt med sundhedsvæsnet, som kan undersøges. Det skal ses som et værktøj, der kan facilitere anvendelse af data, som genereres i forbindelse med de primære processer i Region Nordjylland, til sekundære formål. Det sekundære formål er i denne kontekst identificering af kemiske biomarkører til detektering af KRC.

Datavarehuset indeholder kun data og det sekundære formål indebærer at omdanne individuelle, anonymiserede patient data til generel information. Når data genbruges til sekundære formål, skal data omdannes til information, der kan understøtte formålet. Denne proces er forbundet med problemstillingerne: data integrering og kontekstafhængighed. Begge disse problemstillinger kan løses, men det er ressourcekrævende.

- Omdanne data til information.

    1. Data er kontekst afhængige.
    2. Mange datakilder er enkeltstående informationssystemer uden integrering[40].

Region Nordjyllands datavarehus er et dimensionelt datavarehus, hvor klassificerede data fra både PAS og LABKAII er implementeret i en dimensionel datamodel. Denne opbygning gør det muligt at integrere diagnosekoder med kemiske analyser. Klassificerede data betyder, at data

er udredt fra den primære kontekst, så data er sammenlignelige. Det betyder, at statistiske dataanalyse bliver mulig.
Datavarehuset løser dog ikke disse problemstillinger ubetinget. Både integrering og kontekstafhængighed skal overvejes grundigt i metoden, når data mining skal bruges til at finde sammenhænge i data.

Med disse forbehold omkring anvendelse af data til sekundære formål, er datavarehuset en datakilde, der faciliterer integrering af klassificerede patientadministrative data og laboratoriedata. Denne datakilde kan undersøges for potentielle kemiske biomarkører for KRC, hvilket leder frem til problemformuleringen:

Hvordan er det muligt, at identificere kemiske biomarkører til detektering af kolorektal kræft, givet et datavarehus med patientadministrative data og laboratoriedata?

Metode til problemløsning 3

I dette projekt er der fokus på at identificere kemiske biomarkører for kolorektal kræft (KRC) vha. de laboratoriedata og patientadministrative data, der ligger i Region Nordjyllands (RN) datavarehus. Dette lægger op til anvendelse af principper og metoder inden for beslutningsstøtte herunder data mining. Dette kapitel vil beskrive metoden, hvorved dette blev realiseret, samt hvilke valg, der er taget, og overvejelser, der er gjort for, at nå frem til metoden. Først præsenteres overvejelser omkring studiedesign, derefter præsenteres metoden til data mining.

3.1 Overvejelser omkring studiedesign

Som beskrevet ovenfor ligger dette projekt op til anvendelse af metoder inden for beslutningsstøtte. Beslutningsstøtte er et meget bredt emne, som omfatter det at gøre relevant information om en proces eller en patient, synlig for beslutningstageren[51].
Beslutningsstøtte kan inddeles i to kategorier afhængigt af hvilken viden, der anvendes til at yde beslutningsstøtte. Den første type bygger på viden fra litteraturen, og beslutningsstøtten kaldes da *vidensbaseret. Vidensbaseret* beslutningsstøtte kan f.eks. være en database, der indeholder viden om diagnoser og deres symptomer. Dertil hører en beslutningsmekanisme, som mapper patients data til de diagnoser, som passer på patientens data. Dermed gøres beslutningsstøttende information synlig for beslutningstageren [52].
Den anden type beslutningsstøtte bygger på maskinlæring og ikke på viden fra litteraturen. Denne type kaldes *ikke-vidensbaseret* og bruges ofte når litteraturen ikke præsenterer evidensbaseret viden ift. beslutningsstøttens formål. *Ikke-videnbaseret* beslutningsstøtte udnytter data mining til at genkende mønstre i data, der enten kan detektere eller prædiktere patientens tilstand. Disse mønstre kan derefter, bruges til at træne algoritmer. Når en algoritme er trænet og valideret, kaldes den for en klassifikator. Klassifikatoren kan bruges til at analysere data tilhørende den enkelte patient og præsentere resultatet for beslutningstagerne [52].
I dette projekt holdes fokus på *ikke-vidensbaseret* beslutningsstøtte, fordi litteraturen ikke præsenterer viden, der kan bruges til at detektere KRC vha. eksisterende laboratoriedata og patientadministrative data [3, 21, 4]. Projektets metode vil ikke fokusere på at udvikle beslutningsstøtte til datavarehuset. Fokus vil i stedet være på data mining i datavarehuset for at finde sammenhænge, der senere kan bruges til beslutningsstøtte.

Da der er behov for at finde ukendte mønstre i data ved at udvinde og teste faktorer fra den store datamængde, kan dette projekt karakteriseres som et data mining projekt. Data mining bruges

ofte til at klassificere sygdomme, finde beslutningsstøttende regler til den kliniske praksis, samt til at finde ny viden [53, kap. 1].

Dette projekt udnytter data, der er blevet opsamlet til et andet formål. Det er vigtigt at bemærke, at data allerede er opsamlet inden projektets start, og dermed er udfaldet kendt. Altså patienterne har allerede fået stillet kræftdiagnosen. Da data allerede er opsamlet før projekts start, klassificeres dette projekt som et retrospektivt studie.
Ved retrospektive studier er der både fordele og ulemper, man skal være opmærksom på. Fordelen er, at data er tilgængelige med det samme, der skal ikke bruges ressourcer på dataopsamling. I retrospektive studier kan et større antal patienter og et større antal faktorer inkluderes uden, at omkostningerne af studiet øges.
I retrospektive studier kan det være svært at tage højde for *confounding factors*, fordi data er indsamlet til et andet formål, før studiet er designet[54]. Konceptet *confounding factors* dækker over sammenhænge i data, som er tilstede pga. af en helt anden faktor end KRC. F.eks. kan data mining afsløre, at hypertension er god til at adskille patienter med KRC fra en kontrolgruppe, hvis alle patienterne med KRC ved et tilfælde også har hypertension, og kontrolgruppen ikke har. Hvis en sådan algoritme overføres til et andet datasæt, så falder ydeevnen drastisk, da hypertension ikke har noget med KRC at gøre. Samtidig betyder den retrospektive karakter af projektet, at det er begrænset til de data, der allerede er opsamlet. Som forsker, der bruger retrospektive data, kan man ikke stille krav til dataopsamlingen, fordi opsamlingen allerede er overstået, ved studiets start[55, s. 1131-3].
Ulemperne ved retrospektive studier betyder, at resultatet sjældent kan bruges til at drage definitive konklusioner. Retrospektive studier er derimod rigtig gode til at finde og formulere hypoteser, som senere kan bruges til at danne fundamentet for stærkere prospektive studier med potentiale til at drage konkrete konklusioner.

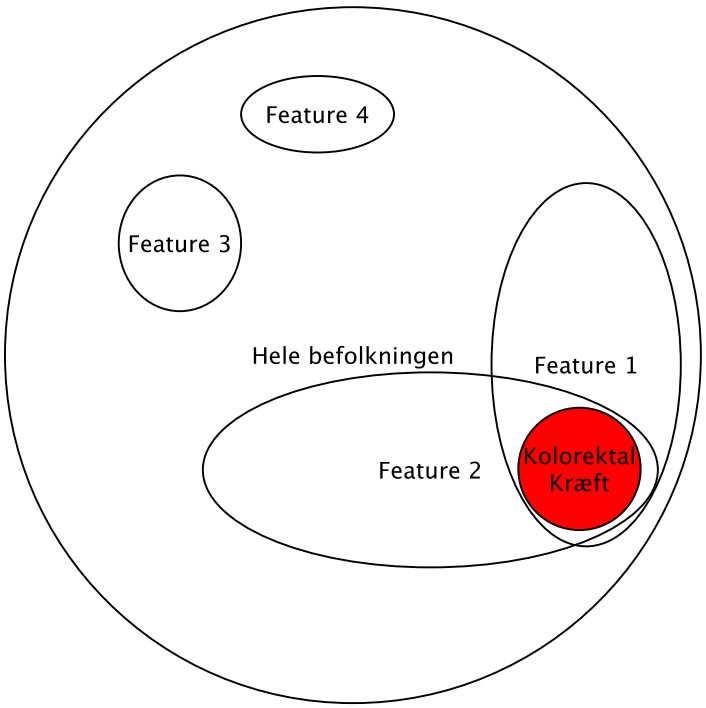
Det er også vigtigt at overveje konteksten af projektet. Dette projektet skal ses i kontekst med screening, fordi formålet er identificering af biomarkører til detektering af KRC, som kan facilitere en tidligere indsats. I screening er man interesseret i at have en meget høj sensitivitet på bekostning af specificiteten, fordi en høj sensitivitet betyder, at ingen patienter med KRC overses. På bekostning af specificiteten betyder også, at der er behov for opfølgende undersøgelser af de patienter, der bliver udvalgt som værende syge af testen med den høje sensitivitet [56]. Det at gå på kompromis med specificiteten tillades kun, fordi der er andre metoder så som koloskopi, der kan bruges til at stille den endelige diagnose. Detekteringen eller screeningen skal nedbringe antallet af patienter, som skal igennem f.eks. koloskopi.
Screeningskonteksten skal ses i forhold til en diagnosticeringskontekst, hvor der er stor interesse i, at have en meget høj specificitet, for at undgå at administrere behandling til patienter, som ikke har sygdommen.

3.2 Metode til dataanalyse - et mønstergenkendelsessystem

Metoden til dataanalyse i dette projekt vil basere sig på mønstergenkendelse, der kan anvendes i forbindelse med datamining. Anvendelsen af mønstergenkendelse stemmer godt overens med målet for dette projekt. Detektering er tæt relateret til det at kunne klassificere patienter [57].
Datamining anvender mønstergenkendelsealgoritmer til at finde sammenhænge i data. Mønstergenkendelse er et forskningsområde, der studerer udviklingen af systemer, der

genkender mønstre i data. Formålet med mønstergenkendelse er at finde og klassificere specifikke grupper i data [58, Kap. 1], hvilket understøtter projektets formål.

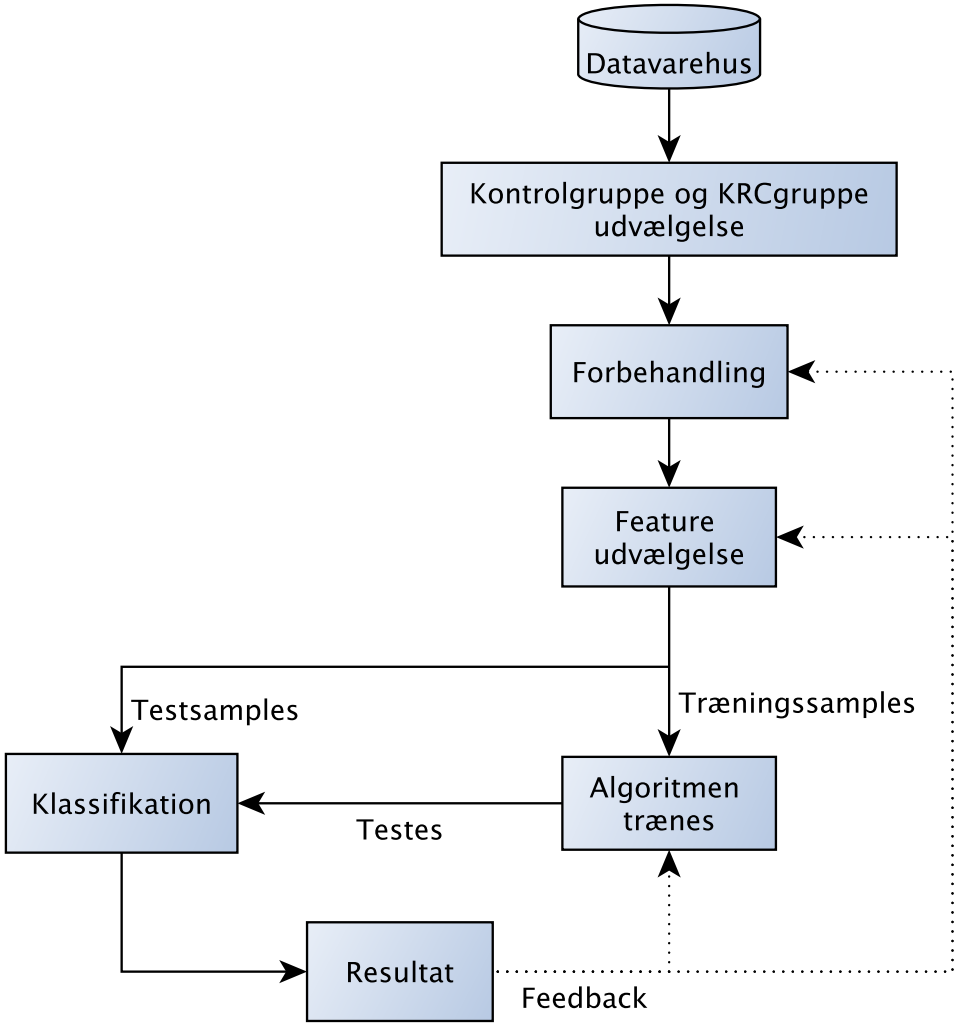
Mønstergenkendelse giver mulighed for at undersøge flere hundrede faktorer ad gangen. I mønstergenkendelsesterminologien kaldes faktorer for features. F.eks. kan man lave en to-dimensionel undersøgelse, hvor features kombineres to ad gangen indtil alle features er testet. Dermed understøttes den multivariate karakter, som anbefales i problemanalysen, ved at kombinere biomarkører til et panel.
Venn-diagrammet på figur 3.1 giver en simplificeret illustration af mønstergenkendelsen. Venn-diagrammet viser hele befolkningen, hvor en del af denne befolkning har kolorektal kræft (KRC). Hos alle med KRC kan *Feature 1* observeres, men *Feature 1* kan også observeres hos mange andre personer i befolkningen. Det samme kan siges om *Feature 2*. *Feature 3* og *Feature 4* overlapper ikke KRC og vil derfor ikke kunne bruges til at klassificerer patienter med KRC. Med mønstergenkendelse testes alle kombinationer af de fire forskellige features. Dermed vil mønstergenkendelsen kunne identificere overlappet mellem *Feature 1*, *Feature 2* og KRC. En af fordelene ved mønstergenkendelse, er at den tillader test af lige så mange features, som det er muligt at inkludere, ikke kun de features, der umiddelbart mistænkes.



Figur 3.1: *Dette venn-diagram illustrerer, hvordan kombinationen af 2 features i teorien kan bruges til at identificere patienter med KRC, hvor en enkelt feature ikke alene kan bruges.*

Givet at der er en forskel på patienter med KRC og patienter uden en kræftdiagnose, siges de to grupper at have forskellige modeller [58, Kap. 1]. Målet med mønstergenkendelsessystemet er at træne en algoritme, der kan adskille disse to modeller. Dermed bliver algoritmen en klassifikator. Det er umiddelbart ikke et stort problem at definere en model for hver gruppe på baggrund af data, så længe de to grupper, der skal adskilles, er velkendte. Udfordringen

består i at bruge disse velkendte grupper, til at udvikle en generaliseret klassifikator, der kan adskille disse modeller, når fremtidige patienter skal klassificeres i de to forskellige grupper. For at teste hvor god den genereliserede klassifikator er, opdeles datasættet i et træningsdatasæt og et valideringsdatasæt.



Figur 3.2: *Illustrer komponenterne i mønstergenkendelsesprocessen, herunder hvordan mønstergenkendelsesalgoritmen trænes og valideres.*

På figur 3.2 ses en illustration af hvordan hele mønstergenkendelsesprocessen er i dette projekt. Først og fremmest udtrækkes data fra datavarehuset til mønstergenkendelses systemet. Dernæst forbehandles data før features udvælges. Når de forskellige features er valgt, vælges samples tilknyttet patientgruppen med KRC og samples tilknyttet patientgruppen uden kræft (kontrol). En lige stor procentdel af hver af disse grupper udvælges som valideringssamples. De resterende data bruges til at træne algoritmen. Når algoritmen er trænet, bruges den til at klassificere valideringssamples for dermed at evaluere klassifikatorens ydeevne. Dette resultat kan bruges til feedback til alle dele af processen.

3.2.1 Udtræk fra datavarehuset

Dette afsnit vil omhandle det benyttede datavarehus samt udtræk af data fra datavarehuset til mønstergenkendelsesalgoritmerne.

På nuværende tidspunkt indeholder FaktaCenteret, som nævnt i problemanalysen, data fra PAS og LABKAII. LABKAII systemet blev iværksat i oktober 2006 i Region Nordjylland, hvilket også markerer datoen for de første laboratoriedata i FaktaCenteret. Tabel 3.1 viser nogle andre datoer, men de tal repræsenterer registreringsfejl i kildesystemet. LABKAII indeholder klinisk biokemiske laboratoriedata. PAS omfatter bl.a. data på kontakter, procedurer og diagnoser. Nøgletal for PAS kan også ses i tabel 3.1.

Det Patient Administrative System (PAS) i Region Nordjylland (RN)	
Antal patienter i PAS	1.098.988
Antal kontakter i PAS	11.256.980
Antal diagnosetyper	35.057
Antal diagnoser stillet	22.057.845
Antal proceduretyper	34.030
Antal procedurer udført	28.114.893
Første henvisningsdato	11/01-1965*
Sidste henvisningsdato	11/12-2013 (Forventet afslutning på henvisning)
LABKAII i Region Nordjylland	
Antal patienter	490.627
Antal patienter med kolorektal kræft	9.915
Antal patient uden kræftdiagnose	391.679
Antal rekvisitioner	7.582.376
Antal prøvesvar	97.566.250
Antal numeriske prøvesvar	73.787.211
Antal analysetyper	2.246
Antal leverandører	565
Dato for første prøvesvar	20/08-1997**
Dato for sidste prøvesvar	03/04-2013 (Prøve bestilt til dato)

Tabel 3.1: *Nøgletal fra FaktaCenteret. * Fejl i kildesystem, formentlig omkring 1970. ** Fejl i kildesystem, bør være 01-10-2006.*

I udtrækket fra datavarehuset var der to typer af patienter, som skulle identificeres. Patienter med kolorektal kræft (KRC) blev samlet i KRC-gruppen og patienter uden en kræftdiagnose blev samlet i kontrolgruppen. Alle inkluderede patienter var over 18 år i den inkluderede periode.
Mht. til kontrolgruppen kan andre kræftdiagnoser påvirke resultaterne af dataanalysen med et forhøjet antal falske positive resultater, da algoritmen måske også identificerer andre typer kræft end KRC. En metode, der kan finde flere typer kræft, er ikke en dårlig ting, men inkludering af patienter med andre typer kræft som kontrol vil påvirke resultatet negativt.
Man kan argumentere for, at andre diagnoser også vil påvirke resultatet ligesom andre kræftdiagnoser, men er metoden ikke robust overfor netop alle disse andre diagnoser, kan den ikke bruges i klinisk praksis. Ved at ekskludere alle de andre kræftdiagnoser nedsættes risikoen for falske positive pga. outliers i data fra kontrolgruppen. Omvendt ved at vælge en kontrolgruppe, som minder om den gruppe patienter, der ses i klinisk praksis, øges styrken af dataanalysen, men på bekostning af præcisionen.

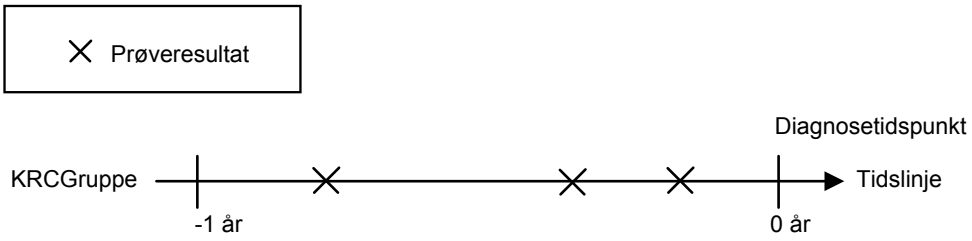
Det næste trin var at udtrække prøveresultater fra datavarehuset, som tilhørte de inkluderede patienter. Dette projekt er interesseret i indikatorer, som kan bruges til at detektere KRC, før diagnosen er blevet stillet og registreret i PAS. Derfor var det nødvendigt, at identificere prøveresultater fra før diagnosen blev stillet. Et studie i screening af KRC udført i Minnesota viste, at screening for okkult blødning hvert år sænker mortaliteten med op imod 33% ift. ingen screening[59]. Et andet studie viste også en 6% lavere mortalitet med screening hvert år ift. screening hvert andet år[19]. Resultaterne af disse undersøgelser viste, at der på et år for nogle tilfælde af KRC kan ske ændringer, som kan fører til en øget mortalitet. Derfor blev det valgt i dette projekt at inkludere prøveresultater, der var leveret op til et år før diagnosen. Dvs. at prøveresultaterne i datavarehuset tilhørende patienter med KRC afgrænses imellem et år før henvisningsdatoen og og henvisningsdatoen. Henvisningdatoen er tilknyttet kontakten, hvor diagnosen KRC blev stillet. Nedenstående tabel 3.2 viser diagnosekomplekset for KRC. Et eksempel på etårsintervallet for KRC-gruppen kan ses på figur 3.3. Det var kun den første kontakt, hvor aktionsdiagnosen KRC var til stede, der blev valgt. Dvs. at der kun blev udvalgt én sample pr. patient.

SKS-Diagnosekomplekset for kolorektal kræft
◇ [DC00-DD48] Neoplasmer
• [DC00-DC96] Kræftsygdomme
◦ [DC15-DC26] Kræft i fordøjelsesorganer
- [DC18] Kræft i tyktarmen
- [DC20] Kræft i endetarmen

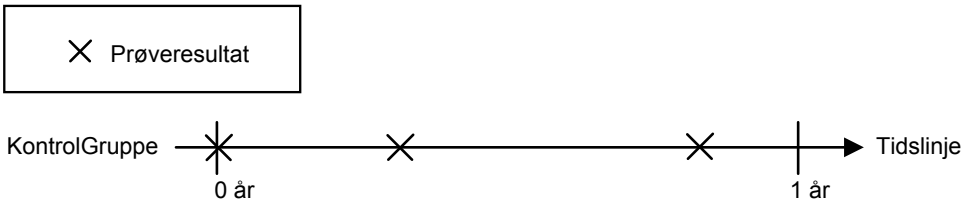
Tabel 3.2: *Diagnosekomplekset for kolorektal kræft. [SKSKode].*

For kontrolgruppen, som bestod af patienter uden kræft, blev der afgrænset imellem datoen for første prøveresultat i perioden fra 01-10-2006 til 01-02-2012 og et år frem. Figur 3.4 illustrer etårsintervallet for kontrolgruppen. I kontrolgruppen blev alle diagnoser ekskluderet, som falder indenfor klassen: neoplasmer (DC00-DD48). Det kan ses i diagnosekomplekset jvf. tabel 3.2.
Kontrolgruppen blev dannet udelukkende på baggrund datoer for prøveresultater, og dermed var der ingen kontaktdatoer involveret, da der i LABKAII kan være mange patienter, som ikke har en kontakt i PAS, fordi patienterne har fået taget prøver hos deres praktiserende læge, uden at have være i kontakt med sygehusvæsnet.
Det blev dog kontrolleret, at den inkluderede patient ikke var tilknyttet en kræftdiagnose. En patient i kontrolgruppen kan i teorien repræsentere flere samples i kontrolgruppen. Dette blev fravalgt, da enkelte patienter kan være repræsenteret mange gange, og dermed skævvride resultatet. Skævvridning af resultaterne sker hvis en patient, som bliver til mange samples, enten er meget nem eller meget svær at klassificere. For kontrolgruppen var én patient derfor også ensbetydende med én sample.

KRC-gruppen indeholdte 9.915 patienter, og kontrolgruppen indeholdte 391.679 patienter på dette stadie. Antallet af inkluderede patient var ikke endeligt på dette tidspunkt i designprocessen af mønstergenkendelsesmetoden, fordi udvælgelsen af features havde stor indflydelse på det endelige antal inkluderede patienter. Alle patienter, som blev inkluderet skulle have data fra de features, som blev inkluderet, og det var langtfra alle, som havde fået taget de sammen analyser.



Figur 3.3: *Figuren viser, hvordan samples findes for KRC-gruppen. Her findes prøveresultater, der er taget op til et år før diagnosetidspunktet.*



Figur 3.4: *Figuren viser, hvordan samples findes for Kontrolgruppen. Her findes prøveresultater, der er taget op til et år efter det første registrerede prøveresultat.*

Ovenstående er et eksempel på, at udtræk af datasæt fra datavarehuset skal gøres med omtanke. Datavarehuset faciliterer kun integrering af de data det indeholder, selve formålet med dataanalysen bestemmer integreringen. Dette nævnes i afsnit 2.4. Dvs. at integrering skal give mening ift. kontekst og formål. Eksempelvis blev ovenstående udtræk af datasættet til både KRC-gruppen og kontrolgruppen påvirket af dataanalysen formål. PAS data og laboratoriedata blev integreret ift. både en periode og en diagnosekode, der begge bestemte inklusion og eksklusion af forsøgspersoner.

3.2.2 Forbehandling

Forbehandlingen af data omfatter normalisering af data. Dette afsnit vil behandle normaliseringsproblematikken i dette projekt samt løsning.

Prøveresultaterne, der er tilknyttet patienterne i de to grupper, kommer fra forskellige analyseapparater. Hvert apparat leverer ikke resultater ift. en standardiseret skala. Dvs. at det er problematisk at sammenligne resultater på tværs af apparaterne, og dermed bliver det svært at udnytte det totale datasæt. Det er især et problem, hvis apparaterne har forskellige producenter, der anvender forskellige biokemiske analysemetoder. Det er et eksempel på, at laboratoriedata ikke kan fortolkes uden viden om laboratoriets normalværdier [24]. En måde at løse denne problemstilling er ved normalisering af resultaterne. Behovet for normalisering af data er et eksempel på, at data stadig har en kontekst, selvom de er klassificeret. Dette var endnu en af de problemstillinger som blev præsenteret i kapitel 2.4, og som skal overvejes i metoden.

Normalisering af analyseresultaterne er et svært område, hvor der skal tages højde for

mange ting. Et apparat kan stå et bestemt sted og hovedsageligt kun levere analyser til en bestemt type patienter. Dermed vil en normalisering blive skævvredet ift. til normaliseringen af et apparat, der bliver brugt til alle patienttyper. Man kan forsøge at kontrollere, om fordelingen af diagnoser tilknyttet analyserne er ens for alle apparaterne, men der findes ingen metoder til at håndtere den slags multivariate sammenligninger af fordelinger. Et eksempel på normaliseringsproblematikken er et apparat, som kun analyserer materiale fra patienter på en intensivafdeling, og et apparat, der kun leverer analyser på patienter fra praktiserende læger. Her vil der nødvendigvis væres forskel på middelværdier og varians på nogle af analyserne, da det må antages, at patienterne på intensivafdelingen er mere syge end patienterne ved den praktiserende læge. Dette er et problem, fordi middelværdier og standardafvigelsen ofte bruges til normalisering.

En måde at komme ud over problemet, er ved selektivt kun at udvælge ét apparat og alle de analyser dette apparat har leveret. Gøres det vil man dog tabe en stor mængde data. En alternativ tilgang, hvor styrken kan øges, er ved at vælge to eller flere apparater, der leverer samme type data og lave sideløbende mønstergenkendelsesanalyser. De sideløbende analyser kan da bruges til at understøtte hinanden, hvormed man får flere uafhængige analyser til at underbygge hinanden. Vælges denne tilgang, introduceres et nyt problem, som er relateret til udvælgelsen af features. Udvælgelsen af features omfatter nogle beregninger (f.eks. lineær regression) på data, som kræver minimum tre prøveresultater pr. analysetype pr. patient. Hvis man kun kan bruge et enkelt apparat, bliver det problematisk af finde nok analyseresultater, fordi det er tilfældigt hvilket apparat, der analyserer prøvematerialet. Dette blev forsøgt med præcis det forventede resultat, at de fleste samples i de to grupper blev tabt. Kun tre af de hyppigst brugte analysetyper (hæmoglobin, leukocytter og P-creatinin) kunne inkluderes før antallet af samples i de to grupper tilsammen faldt til nul.

På baggrund af ovenstående var vi nødsaget til at finde en alternativ tilgang, hvor flere resultater kunne inkluderes. I praksis viste det sig dog, at dette ikke er så stort et problem, som først antaget. Laboratorierne på de enkelte hospitaler forsøger, at holde resultaterne normaliseret internt på hospitalet for at sikre klinikernes fortolkning af resultater. Dermed er resultaterne ikke apparatafhængige, og det giver mulighed for sammenligning og normalisering på tværs af hospitalerne. Med det in mente blev alle analysestyper normaliseret med normalværdier fra det laboratorium, der havde leveret reulstaterne.

I dette projekt blev normalisering udført med koefficienterne middelværdi og standardafvigelse, som blev beregnet ud fra alle data i LABKAII grupperet efter analysetype. Middelværdien blev udregnet med følgende formel:

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n} \tag{3.1}$$

og standard afvigelsen blev udregnet med formlen:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}} \tag{3.2}$$

hvor n er antallet eksisterende numeriske resultater. Alle prøveresultater i kontrolgruppen og KRC-gruppen blev korrigeret med følgende formel:

$$\text{NormaliseretResultat} = \frac{X - \mu}{\sigma} \tag{3.3}$$

3.2.3 Udvælgelse af features

Udvælgelse af features omhandler udvælgelsen af de analysetyper, der skal undersøges for potentiale som biomarkører for kolorektal kræft (KRC), samt hvilke statistiske karakteristika, der skal beregnes for analysetyperne.

Dette projekt undersøgte data, fra før diagnosen KRC blev stillet. Mængden af data tilgængelig i den periode kan risikere at være meget lille, fordi det er før den intensive kontakt med sundhedsvæsnet, hvor diagnosen blev stillet. Det var dog vores formodning at patienten har været igennem et udredningsforløb for KRC eller behandlingsforløb for anden sygdom inden diagnosen blev stillet, hvor der også blev taget blodprøver. Det er disse data, vi var interesseret i at undersøge for potentielle indikatorer, og det satte en begrænsning ift. mængden af tilgængelige data.

Tidligere i afsnittet om forbehandling blev et krav om minimum tre prøveresultater pr. analysetype introduceret. Det hænger sammen med interessen i at beskrive både ændringer og niveauforskelle i de prøveresultater, der er indenfor den udvalgte periode hos en given patient. F.eks. kan man forestille sig, at en patients prøveværdier falder eller stiger i perioden. Et sådan fald eller stigning vil kunne beskrives med en lineær regression. Man kan også forestille sig, at patientens tilstand bliver mere og mere ustabil. Stabilitet eller ustabilitet kan komme til udtryk ved standardafvigelse. Man kan også forestille sig, at en ændring allerede var sket hos en patient med KRC inden den udvalgte periode. Så vil data i perioden udtrykke et bestemt niveau. Er det tilfældet, kan niveauforskellen undersøges hos de to patientgrupper med middelværdien.

I både KRC-gruppen og kontrolgruppen blev der beregnet, middelværdi, standardafvigelse og lineær regression for hver inkluderet analysetype for hver enkelt patient. Beregningen af lineær regression har behov for minimum tre prøveresultater, deraf kom kravet om minimum tre prøveresultater pr. analysetype. pr. patient, for at patienten kunne inkluderes. For at undgå fænomenet *missing features*, så blev en patient kun inkluderet, hvis der var minimum tre prøveresultater fra hver af de inkluderede analysetyper. Listen herunder opsummerer, hvad de forskellige statistiske metoder evaluerer.

- Lineær regression: Evaluerer fald eller stigning i data.
- Middelværdi: Evaluerer niveauforskelle i data.
- Standard afvigelse: Evaluerer ustabilitet i data.

Det næste skridt var at vælge de analysetyper, som skulle undersøges. Her var det nødvendigt at identificere og vælge de analysetyper, som oftest bliver bestilt til andre formål. Dette gjorde, at antallet af samples i de to grupper forblev så højt som muligt. Var analysetyper, der sjældent bruges, blevet valgt, så ville antallet af tilgængelige samples falde. I et screeningsperspektiv er det også en god egenskab at udnytte de oftest bestilte analysetyper, fordi data er til rådighed.
For både kontrolgruppen og KRC-gruppen blev datavarehuset brugt til at udtrækket analysetyperne. Listen kan ses i tabel 3.3. Tabellen viser de 15 mest benyttede analysetyper for begge grupper. Dette er også de 15 analysetyper, som det var muligt at benytte i dette studie uden at tabe for mange patienter. Dette skyldtes, at der ikke var en stor nok gruppe af patienterne, som havde minimum tre prøveresultater i flere end de 15 mest anvendte analysetyper. Ved inkludering af den 16'ende mest anvendte analysetype (*Pt-estimeret GFR pr. 1,73 m^2 (eGFR)*), da faldt antallet af patienter i kontrolgruppen fra 6.149 til 232. Dette blev vurderet som et for stort tab.

	Analysenavn
1.	B-Hæmoglobin
2.	P-Creatinin
3.	P-Kalium-ion
4.	P-Natrium-ion
5.	B-Leukocytter
6.	P-Albumin
7.	P-C-reaktivt protein (CRP)
8.	P-Calcium(total)
9.	P-Calcium(total)(alb.korr.)
10.	B-Thrombocytter
11.	P-Koagulationsfaktorer 2,7,10 (INR)
12.	P-Basisk phosphatase(BASP)
13.	P-Alanintransaminase(ALAT)
14.	P-Bilirubiner
15.	B-Erythrocytter(MCV)

Tabel 3.3: *De 15 mest benyttede analysetyper i FaktaCenteret*

I alt blev 9.915 patienter med diagnosen KRC identificeret i RN's PAS i perioden fra 2006 til 2013. Kun 125 af disse havde nok prøveresultater til at blive inkluderet i dette studie. Det understøtter mistanken om at datatætheden i datavarehuset er meget lav før kontakten, hvor diagnosen stilles. Ifølge tabel 3.1 havde kontrolgruppen potentiale til at indeholde 391.679 patienter. Men kun 6.149 havde minimum tre prøveresultater for hver analysetype.

Tabel 4.1 på side 42 i resultatkapitlet viser demografi for de to patientgrupper. Tabellen kan bruges til at danne forventninger om hvilke analysetyper, der har det største potentiale som indikatorere.
I alt blev 15 analysetyper identificeret. I tabel 4.1 på side 42 vises middelværdierne for de 15 analysetyper. Der blev yderligere beregnet en P-værdi vha. *Students T-test*, for at se om analysetyperne hver for sig og en ad gangen kan detektere en forskel i de to grupper. Det kan måske også give en ide om hvilke analysetyper, der vil klarer sig bedst i den senere dataanalyse. For hver analysetype blev 3 statistiske karakteristika beregnet. Dvs. $3 \cdot 15 = 45$ features. Derudover blev alderen også inkluderet, som en feature, pga. af forskellen, der ses i tabel 4.1, og viden, om at 99% af patienter med KRC er over 40 år jvf. A.2. Dvs. at der i alt blev udvalgt 46 features.

3.2.4 Valg af mønstergenkendelsesmetode til klassifikation

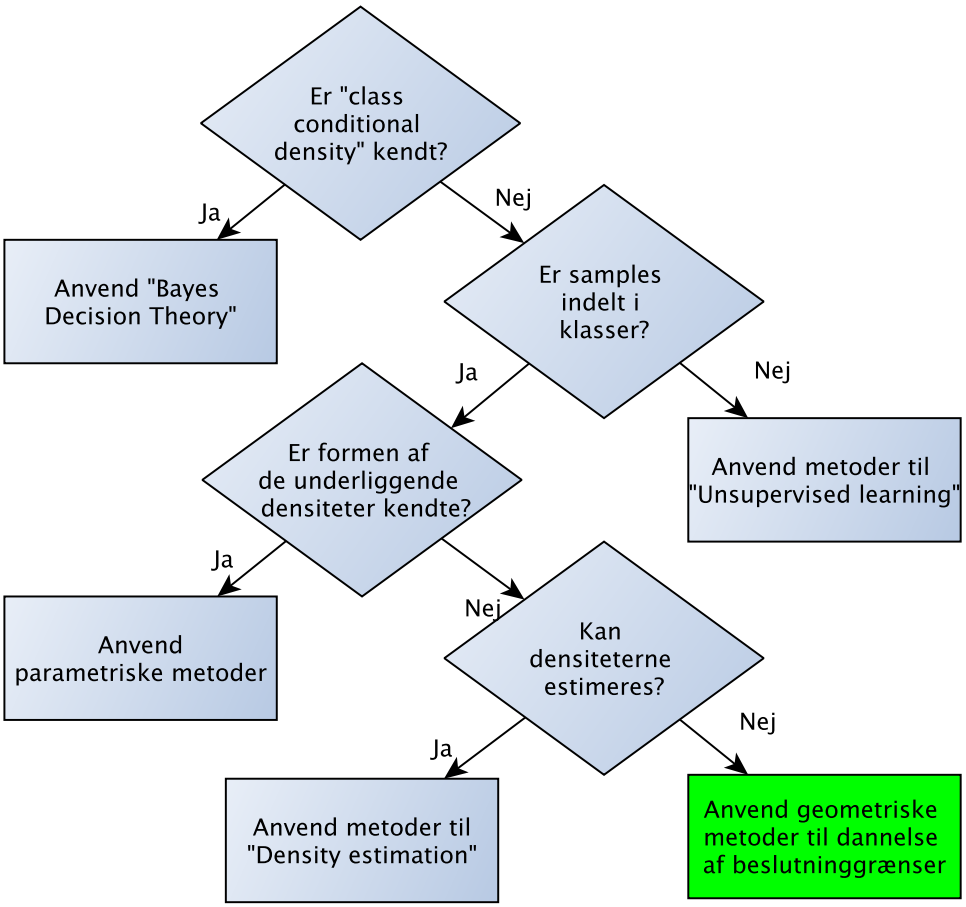
I mønstergenkendelse er der mange forskellige algoritmer, som kan anvendes. Algoritmerne bliver også kaldet klassifikatorer, når de er blevet trænet. Der er fordele og ulemper ved de forskellige klassifikationsmetoder, og der kræves kendskab til datasættet samt et klart mål for dataanalysen, for at kunne vælge den rigtige metode.

Figur 3.5 viser, hvordan typen af klassifikationsmetoder er blevet valgt i dette projekt. I dette projekt er en geometrisk metode blevet valgt. For at komme frem til dette tages stilling til en række spørgsmål, som kan ses i figur 3.5.

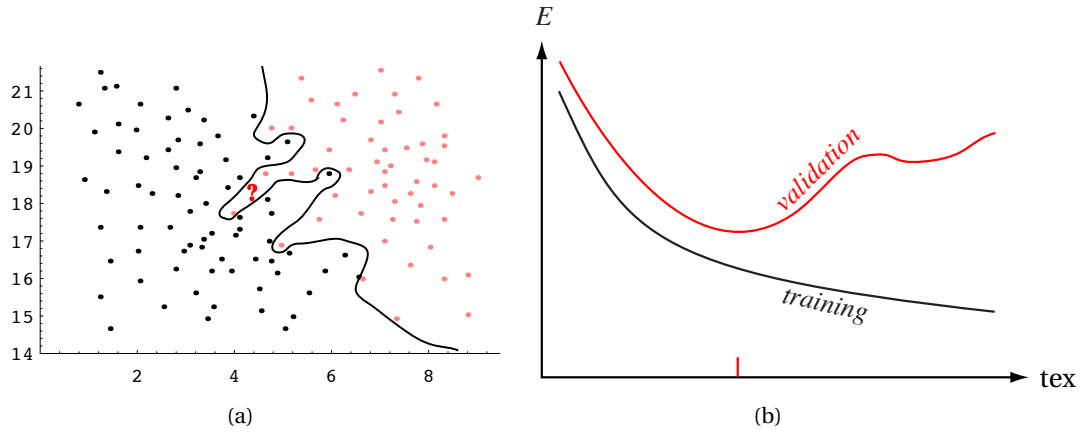
Det første spørgmål der stilles, når der skal vælges klassifikationsmetoder, er: Er *Class conditional density* kendt? *Class conditional density* beskriver sandsyndligheden for at en sample tilhører en bestemt klasse givet en bestemt featureværdi. F.eks. hvad er sandsynligheden for, at en patient har kolorektal kræft (KRC), hvis *B-Hæmoglobin* er under 7,0? Dette er umiddelbart ikke komplekst, men i dette projekt undersøges 46 forskellige features i kombination. At kende *Class conditional density* for en enkelt feature er ikke nok. Den skal kendes for alle kombinationer af features på en gang. Den slags information var ikke til rådighed i dette projekt, hvilket også er et kendt problem i mange andre domæner[58]. Dermed var svaret: **Nej *"Class conditional density"* er ikke kendt**.
Det næste spørgsmål i beslutningstræet relateres til hvilken type information, der er til rådighed om de inkluderede samples. *"Er samples inddelt i klasser?"* Dvs. er hver sample markeret som en kontrolsample eller som en KRC-sample. Eftersom samples i dette projekt var opdelt i en kontrolgruppe og en KRC-gruppe, var samples inddelt i klasser og dermed var svaret: **Ja samples er inddelt i klasser**.
Da klasserne for de enkelte samples er kendte, anvendes *"Supervised learning"*. Det betyder, at både antallet af klasser er kendte, og at algoritmen ved, hvilken gruppe en sample tilhører. Oftest er *"Supervised learning"* algoritmer, mere effektive end *"Unsupervised learning"* algoritmer, fordi de har et bedre beslutningsgrundlag. Faren ved at anvende *"Supervised learning"* er overtræning af algoritmen. Når man ved hvilken klasse hver sample tilhører, kan man sagtens lave en algoritme så kompleks, at den kan klassificere alle samples, som algoritmen trænes med, korrekt. Dette fænomen er illustreret på figur 3.6(a). Overtræning skader ofte generaliserbarheden af algoritmen. Testes den trænede algoritme med et nyt sæt af samples, som algoritmen altså ikke er blevet trænet med, så vil resultaterne af denne test, ofte blive meget dårlige hvis algoritmen er overtrænet. Dette er illustreret på figur 3.6(b). Det ses desuden, at desto længere tid man bruger på træning af algoritmen, desto lavere fejlrate vil man opnå, så længe at træningssamples også bruges til at teste. Derimod vil valideringen af algoritmen med testsamples evaluere til en højere fejlprocent. Overtræning undgåes ved at tage en del af KRC-gruppen og en del af kontrolgruppen, som kun bruges til validering og ikke til træning af algoritmen. Dette vil blive uddybet yderligere i næste afsnit om validering.

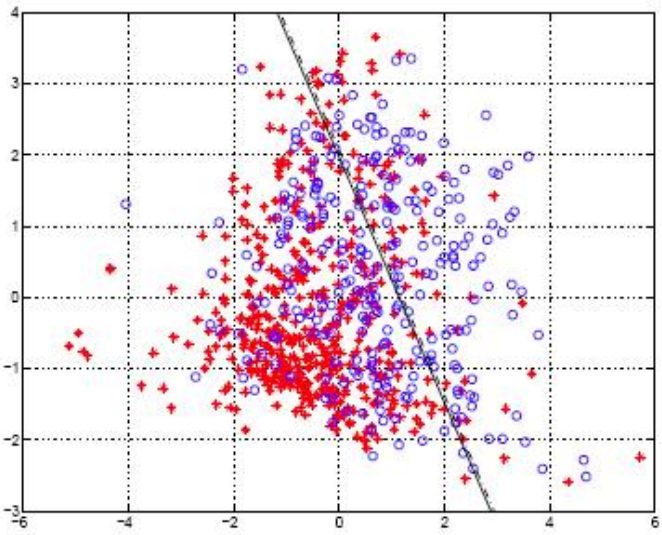
Derefter er der to typer af algoritmer, man kan vælge imellem, *"non-parametriske"* og *"parametriske"*. Det afgørende er, om formen af de underliggende densiteter (*"Class conditional density"*) for de to klasser er velkendte, og at de er normalfordelte. Som nævnt tidligere er der ingen information om *"Class conditional density"* til rådighed. Med hensyn til normalfordeling kan man oftest antage, at biologiske data er normalfordelte[58]. Men det er en antagelse, som er problematisk i praksis, især når de biologiske data kommer fra syge patienter. At være syge er ikke en normal tilstand, og derfor kan det være svært at antage, at data fra disse patienter er normalfordelte. Derfor er svaret: **Nej, formen af de underliggende densisteter er ikke kendte**. Hvis formen af densiteterne ikke er kendt, bruges non-parametriske metoder. Nogle non-parametriske metoder forsøger at estimere formen af *"class conditional density"*. Denne type af metode lider under det fænomen, der kaldes: *"The curse of dimensionality"*. Det betyder, at antallet af samples, som skal bruges til at give en god estimering af formen, stiger eksponentielt med antallet af dimensioner (kombinationer af features). Antallet af samples til rådighed i dette projekt er relativt begrænsede, mens antallet af features er relativt højt. Der er 46 features, som skal undersøges, og 50 features er et moderat til højt antal iflg. [58]. Derfor er svaret: **Densiteterne er uhensigtmæssige af estimer**. Derfor blev der anvendt geometriske metoder i

dette projekt. De geometrisk metoder bruges til at danne en beslutningsgrænse mellem KRC-gruppen og kontrolgruppen. Beslutningsgrænsen kan bl.a. dannes med en diskriminantanalyse, som modellerer forskellen mellem de to klasser.



Figur 3.5: *Dette beslutningstræ modificeret fra [60] viser hvordan en metodetyper i mønstergenkendelsen vælges. Den grøn boks illustrer metodetypen, som anvendes i dette projekt*





Figur 3.6: *De røde punkter og de blå cirkler tilhører hver deres klasse. Den lige linje er en linear beslutningsgrænse.*

Det er nødvendigt manuelt at vælge hvilken form beslutningsgrænsen skal have. I den beslutning er det vigtigt at huske overtræningsprincippet. Derfor blev formen af beslutningsgrænsen i dette projekt valgt til en simpel lineær og en lidt mere kompleks kvadratisk form.
En beslutningsgrænsen dannes ved at beregne en lineær diskriminant funktion eller en kvadratisk diskriminant funktion for hver af de to grupper. Dvs. der beregnes en funktion, der beskriver KRC-gruppen og en funktion, der beskriver kontrolgruppen. Derefter findes forskellen mellem de to funktioner, som netop beskriver beslutningsgrænsen. Følgende formel blev brugt til at beregne den lineære diskriminant funktion:

$$g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + \omega_0 \tag{3.4}$$

x er en input vektor, som har samme længde som antallet af alle features. Dvs. testes to features i kombination altså en to dimensionel test, så indeholder **x** 2 features. **w** kaldes "weight vector"og den bestemmer, hvordan en features skal vægtes, så man undgår skalaringsfejl (F.eks. kunne alderen komme til at vægte højere end B-hæmoglobin, fordi middelværdien af alderen er væsentlig højere end middelværdien for hæmoglobin i.flg. tabel4.1).

Den beregnes for hver feature ved formlen:

$$\mathbf{w} = \Sigma^{-1}\mu \tag{3.5}$$

Hvor Σ er *"covarians matrix"* og μ er en vektor med middelværdier af features beregnet ud fra de samples, som algoritmen trænes med. *Covariansen* fortæller noget om, hvordan feature ændrer sig ift. hinanden. $\mathbf{w}^t \cdot \mathbf{x}$ er et skalarprodukt, som siger noget om hvordan den lineære funktion er orienteret i forhold til samples i gruppen. ω_0 kaldes: "off-set"og beskriver placering af samples i featuresrummet. Den beregnes udfra:

$$\omega_0 = -\frac{1}{2}\mu^t\Sigma^{-1}\mu + ln(P(\omega_i)) \tag{3.6}$$

hvor $P(\omega_i)$ er sandsynlighed for at en sample tilhører gruppen *i*. I dette projekt er der to grupper. $P(\omega_K RC)$ er sandsynligheden for at en patient har KRC ud fra det totale antal samples, som bruges til træning.

Funktionen (3.4) g(x) beregnes for både kontrolgruppen og KRC-gruppen. Da man vil adskille de to grupper, er det forskellen, der har interesse. Derfor skal disse to funktioner omdannes til én funktion, der beskriver beslutningsgrænsen mellem de to grupper. Se figur 3.6. Beslutningsgrænsen dannes ved at trække de to diskriminant funktioner fra hinanden:

$$g_{KRC}(\mathbf{x}) = g_{Kontrol}(\mathbf{x}) \Leftrightarrow g_{KRC}(\mathbf{x}) - g_{Kontrol}(\mathbf{x}) = g_{KRC-Kontrol}(x) = 0 \tag{3.7}$$

Nu er algoritmen trænet og den er blevet til en klassifikator. Når klassifkatoren skal valideres, indsætte én sample i $\mathbf{x}$ ad gangen. Hvis $g_{KRC-Kontrol}(x) > 0$ så beslutter klassifikatoren, at samplen tilhører KRC-gruppen, og hvis $g_{KRC-Kontrol}(x) \leq 0$, så tilhører samplen kontrolgruppen.

Den kvadratiske diskriminant funktion udregnes efter samme principper, der tilføjes kun et led mere. En kvadratisk diskriminant funktion udregnes ved formlen:

$$g(\mathbf{x}) = \mathbf{x}^t\mathbf{W}\mathbf{x} + \mathbf{w}^t\mathbf{x} + \omega_0 \tag{3.8}$$

Hvor $\mathbf{W} = -\frac{1}{2}\Sigma^{-1}$.

Når algoritmen trænes, skal dimensionaliteten af analysen vælges. Dette gøres ved at vælge hvor mange features, der skal kombineres ad gangen. Igen spiller overtræningsprincippet ind, idet at det er usandsynligt, at samtlige features vil kunne bruges til at adskille grupperne. Nogle features vil blot tilføje ekstra kompleksitet uden at forbedre resultatet. Fra mønstergenkendelse i andre domæner end sundhedsektoren, er det kendt, at der er en balance mellem antallet af features og klassifikationsfejl [60]. Målet er først og fremmest, at vælge hvor mange features, der skal kombineres ad gangen, og derefter iterere igennem alle kombination af features for at finde den bedste kombination.
I dette projekt blev der lavet to analyser. I den første analyse, blev en 1-dimensionel test, en 2-dimensionel test og en 3-dimensionel test udført. En 2-dimensionel test kombinerer to features ad gangen indtil at mulige kombinationer er testet. På baggrund af denne første og indledende analyse udvælges de tre bedste kombinationer af features fra de tre tests. På baggrund af de udvalgte features fra den indledende analyse laves en afgrænset analyse, hvor de udvalgte features bruges til at kombinere fire features ad gangen op til kombinationer af ti features ad gangen (ti-dimensionel).
Grunden til at der blev lavet en indledende og en afgrænset analyse, er et spørgsmål om

hvor mange beregninger, der skal laves, og hvor lang tid det tager, at lave beregninger. F.eks. kombineres 46 features 2 og 2 på alle mulige måder, så skal algoritmen trænes og testes $46 * 45/2! = 1.035$ gange. For hver gang at algoritmen trænes og testes, tager det ca. 280 millisekunder [1] . Af tabel 3.4 fremgår det, hvordan analysetid og antallet af iterationer udvikler sig i forhold til antallet af dimensioner, som kombineres. Det ses at analysetiden og antallet af iterationer, stiger eksponentielt når antallet af kombinerede features stiger. Det var derfor nødvendigt at nedsætte antallet af features, som kombineres, for at minimerer den tid det tager at lave analyser.

Med det in mente vil der i resultatafsnittet blive præsenteret resultater fra en 1-,2- og 3-dimensionel analyse af de 46 features. Derefter præsenteres de udvalgte features. Til sidste præsenteres resultaterne for 4-,5-,6-,7-,8-,9- og 10-dimensionelle analyser med de udvalgte features.

Dimensioner	**Antal iterationer**	**Analysetid** [1]
1	46	ca. 10 sek.
2	1.035	ca. 5 min.
3	15.180	ca. 70 min.
4	163.185	ca. 12 timer
5	1.370754	ca. 107 timer

Tabel 3.4: *Tabellen illustrer udviklingen i antal iterationer og analysetid ift. antal features der kombineres ad gangen(Dimensioner).*

3.2.5 Validering af algoritmerne

I dette projekt er der et begrænset antal samples til rådighed. 125 samples i KRC-gruppen og 6.149 i kontrolgruppen. En del af sample-mængden skal bruges til at træne algoritmerne, men der skal også bruges en del til at validere klassifikatorerne.

Der findes flere forskellige metoder til validering: *holdout* og *leave-one-out* er de to metoder, som oftest bruges. *Holdout* deler det totale antal samples i en træningsmængde og en valideringsmængde. Dvs. at algoritmerne ikke trænes med de samme samples, som de valideres med. Det følger også overtræningsprincippet, som ses på figur 3.6(b). Denne metode kræver mange samples, fordi den ikke genbruger samples, men den er meget hurtig, da der kun skal laves én træning og én validering for at få algoritmens ydeevne.
Leave-one-out bruges ofte hvis antallet af samples er meget lavt, fordi den genbruger samples til både træning og evaluering. Den holder én sample tilbage ad gangen til validering, mens algoritmen trænes med de resterende samples. Dette gøres indtil, hver enkelt sample har været holdt tilbage til validering. Hvis man brugte *Leave-one-out* validering i dette projekt, ville valideringen af en enkelt algoritme kræve $6.149 + 125 = 6.274$ (det totale antal samples) træninger og valideringer for at nå det endelige resultat. Det er uholdbart ift. beregningerne om beregningstid lavet i tabel 3.4. Derfor vælges *holdout* metoden i dette projekt.
I litteraturen holdes mellem 10% og 50% af det totale antal samples tilbage til validering [58, 61]. I dette projekt holdes **20%** af antallet af samples tilbage.

[1] Testet med: Apple MacBook Air 13,3"midt-2012, Intel Core i5 1,8Ghz (Ivy Bridge), 8GB DDR3-ram, 256GB SSD, Mac OS-X 10.8.3, Matlab R2012b

Når klassifikatoren er trænet, bliver den brugt til at klassificere de 20% af samples, der er blevet gemt til validering. Klassifikatoren ved, hvilken klasse samples i træningsmængden tilhører, men den ved ikke noget om, hvilken klasse samples i valideringsmængden tilhører. Når klassifikatoren har klassificeret alle valideringssamples, sammenlignes klassificeringerne med den sande gruppe, som samples i valideringsmængden tilhører. Resultatet af dette giver et mål for klassifikatorens evne til at klassificere korrekt. Klassifikatoren leverer resultatet i en *confusion matrix*, som beskriver antallet af sande negative, falske negative, sande positive og falsk positive udfald.

Tabel 3.5 viser en *confusion matrix*. De to koloner til højre er klasserne. De to nederste rækker viser, hvordan algoritmen klassificerer samples. *Sand Positiv* er antallet af samples, der er klassificeret som KRC og som tilhører klassen KRC. *Falsk Positiv* er samples, der klassificeres som KRC, men som tilhører kontrolgruppen. *Falsk Positiv* kaldes også for *type I fejl. Falsk Negativ* er antallet af samples der klassificeres som kontrol, men tilhører KRC-gruppen. *Falsk Negativ* kaldes *type II fejl. Sand Negativ* er antallet af samples, som klassificeres som kontrol og som tilhører kontrolgruppen [62].

	KRC	**Kontrol**
Klassificeret som KRC	Sand Positiv(SP)	Falsk Positiv(FP)
Klassificeret som kontrol	Falsk Negativ(FN)	Sand Negativ(SN)

Tabel 3.5: *Confusion-matrix. "Positiv"og "Negativ"beskriver, hvordan algoritmen klassificerer, mens "Sand"og "Falsk"beskriver om klassificeringen er korrekt.*

Confusion matrix'en bruges til at beregne algoritmens sensitivitet, specificitet, positive prædiktivitet og negative prædiktivitet, som alle fire præsenteres i resultatafsnittet 4.
Sensitivitet er et mål for hvor mange samples, der har kolorektal kræft (KRC) og som klassificeres med KRC. Dermed måles algoritmens evne til at opdage tilfældene af KRC.

$$Sensitivitet = \frac{SP}{SP + FN} \tag{3.9}$$

Specificitet er et mål for antallet af sande negative resultater ift. antallet af samples, som tilhører kontrolgruppen (negativ). Dermed beskriver specificiteten algoritmens evne til at klassificere samples til kontrolgruppen korrekt.

$$Specificitet = \frac{TN}{TN + FP} \tag{3.10}$$

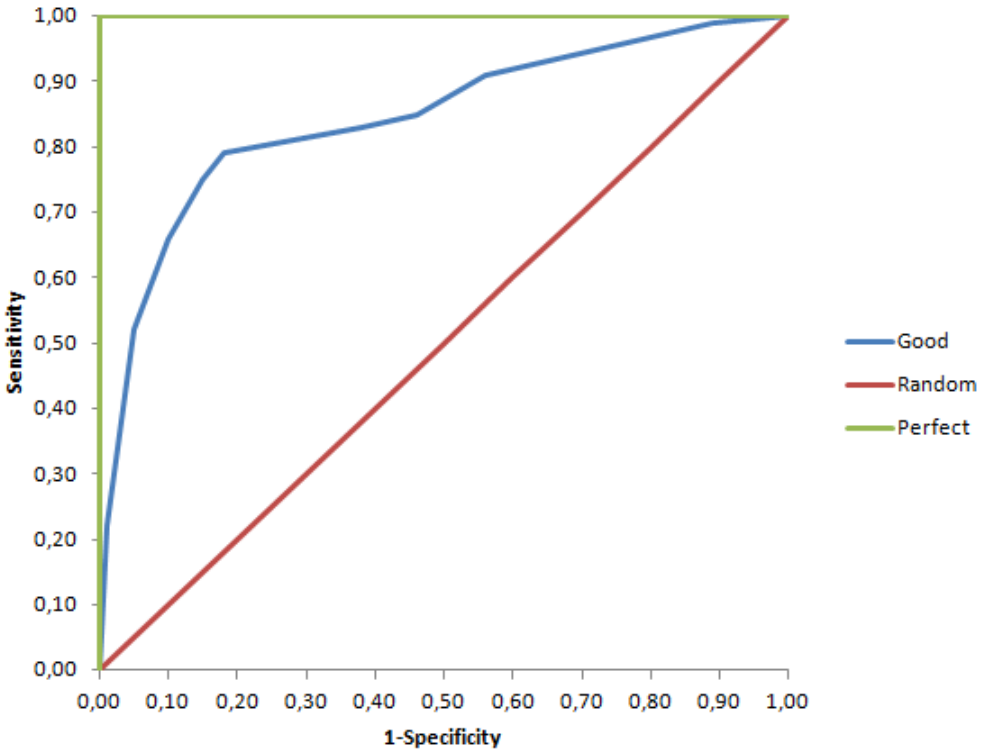
Positiv prædiktivitet er sandsynligheden for at en sample klassificeret til KRC-gruppen også tilhører KRC-gruppen.

$$Positiv\ pred. = \frac{TP}{TP + FP} \tag{3.11}$$

Negativ prædiktivitet er sandsynligheden for at en sample, der bliver klassificeret som tilhørende kontrolgruppen også tilhører kontrolgruppen.

$$Negativ\ pred. = \frac{TN}{TN + FN} \tag{3.12}$$

Da sensitivitet og specificitet afhænger af hinanden kan de vægtes forskelligt. F.eks. er der i et screeningsperspektiv et ønske om at få en højere sensitivitet på bekostning af specificiteten. En screeningsmetode skal have høj sensitivitet, så den ikke overser KRC. Derfor er man villig til at ofre specificiteten.
Når algoritmen trænes, ved algoritmen ikke hvordan sensitivitet og specificitet skal vægtes. Træningen forsøger kun at minimere det totale antal af "falske"resultater. Dvs. summen af "FP"og "FN".
For at finde den optimale sensitivitet ift. specificitet i dette projekt beregnes en *Receiver Operating Characteristic* (ROC) kurve. For at beregne en ROC kurve, manipuleres klassifikatoren i små trin fra en sensitivitet på 0 til en sensitivitet på 1. Dermed kan man vælge den optimale sensitivitet ift. specificitet. På figur 3.7 kan tre ROC kurver ses. Den grønne linje er den optimale, fordi den har 100% sensitivitet og 100% specificitet, men dette er sjældent opnåeligt i praksis [31]. En rigtig god ROC kurve vil ligge tæt på den grønne linje. Den blå ROC kurve er et mere realistisk eksempel, som viser hvordan sensitiviteten påvirker specificiteten. Kan algoritmen ikke kende forskel på klasserne, vil den rød linje opstå. Den røde linje svare til at algoritmen kun har 50% chance for at klassificerer rigtigt. Altså er det tilfældigt hvilken klasse algoritmen tildeler samples.

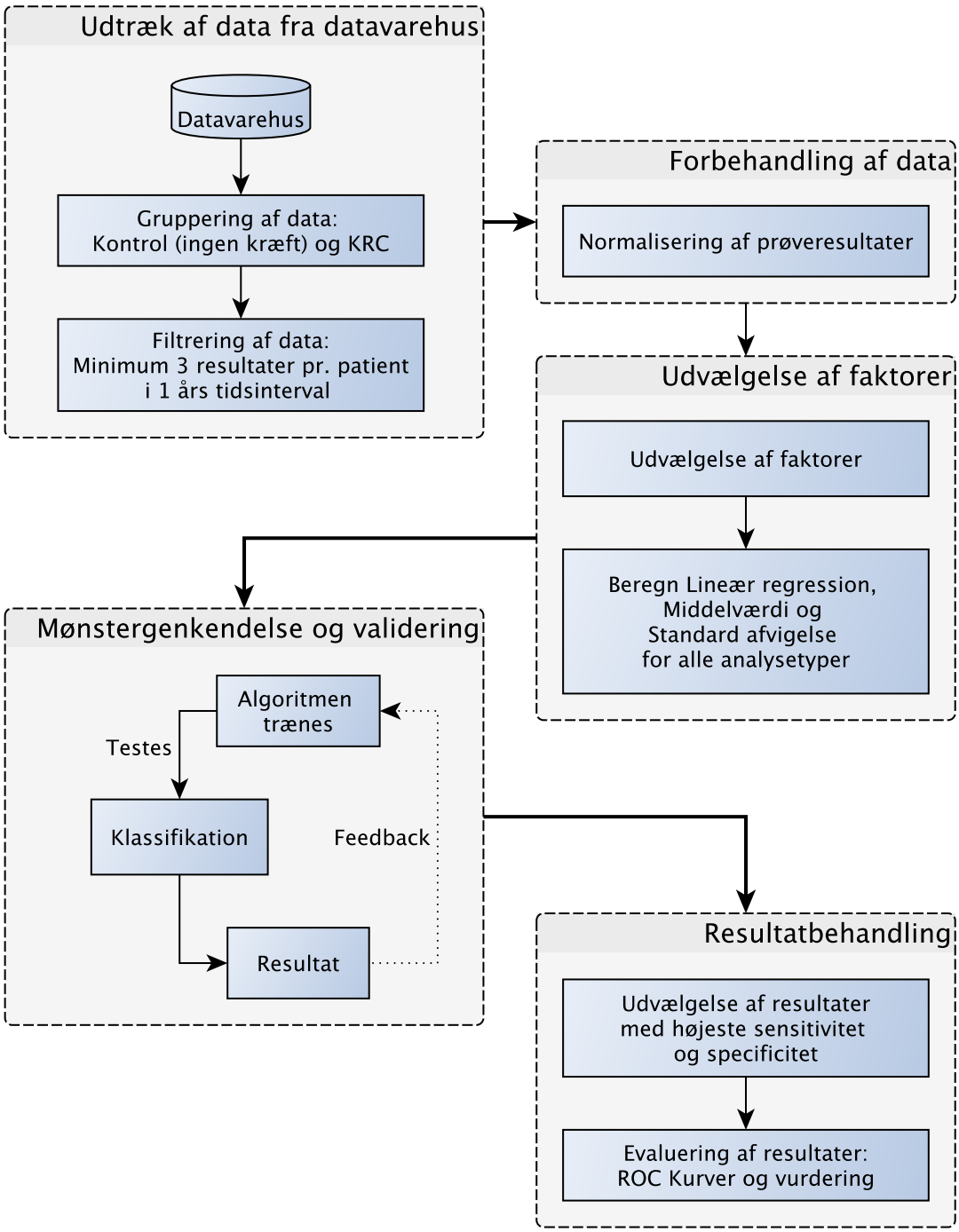


Figur 3.7: *Figuren viser tre ROC kurver. Y-aksen er sensitivitet og x-aksen er 1-specificitet. ROC kurven beskriver afvejningen mellem sensitivitet og specificitet. Man leder altid efter en algoritme, som har både en høj sensitivitet og en høj specificitet, dog kan de vægtes forskelligt.*

3.3 Delkonklusion

Metoden til dataanalyse i et datavarehus er omfattende, og det kræver stor indsigt i både data og de metoder, som skal anvendes. Det er metoden til dataanalysen, hvor det største arbejde med at anvende datavarehuset til detektering af KRC findes.

Metoden til datanalyse kan kort opsummeres som følgende:
Først blev data udtrukket fra datavarehuset til kontrolgruppen, der ikke måtte indeholde patienter med en kræftdiagnose, og data til KRC-gruppen, der indeholder patienter med diagnosen KRC. Inden data blev udtrukket, blev de enkelte analysetyper normaliseret, så data på tværs af laboratorier kunne sammenlignes. Herefter blev data filteret, så patienter, der havde mindre end tre resultater pr. analysetype inden den inkluderede 1-årige periode, blev ekskluderet. Herefter blev de statistiske karakteristika beregnet for de inkluderede analysetyper. Disse udgjorde i alt 46 features. De statistiske karakteristika var: lineær regression, middelværdi og standard afvigelse. Derefter blev en lineær og en kvadratisk diskriminant analyse udført, hvor algoritmerne blev trænet og valideret med *holdout* 20%. Der blev lavet en indledende analyse med 1-3 dimensionelle tests for at nedsætte beregningstiden. Derefter blev de bedste features udvalgt og brugt i 4-10 dimensionelle test. Både sensitivitet, specificitet, positiv prædiktivitet og negativ prædiktivitet blev beregnet og ROC kurver blev plottet. Figur 3.8 illustrerer trinnene i dataanalysemetoden.



Figur 3.8: *Figuren illustrerer de forskellige processer, der er i dataanalysemetoden.*

Resultater 4

I dette afsnit præsenteres de forskellige resultater fra klassifikatorerne samt et udtræk fra datavarehuset, der illustrerer fordelingen af prøveresultaterne for hver analysetype og demografien af KRC-gruppen og kontrolgruppen.

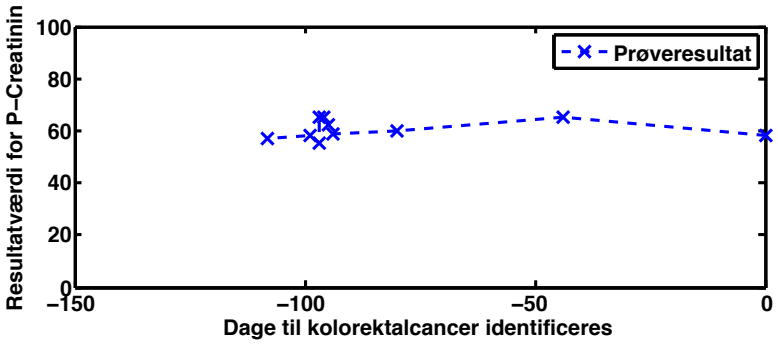
Det skal bemærkes at, resultaterne præsenteret i resultatafsnittet kun er de klassifikatorer, der gav de højeste værdier af sensitivitet og specificitet. Men der var andre featurekombinationer, der kunne bidrage med resultater, som var meget tæt på de bedste klassifikatorer, så er de næst- og tredjebedste klassifikatorer beskrevet i appendiks C.1.

4.1 Fordeling af prøveresultater

Det er tidligere nævnt i metoden, at der for retrospektive data genereret i klinisk praksis ikke er forskningprotokoller for hvilke data, der skal opsamles, og hvordan data skal opsamles, hvornår data skal opsamles og hvor ofte data skal opsamles. Der kan dog være guidelines, som skal følges, men disse guidelines fokuserer på diagnosticering og behandling af patienten, ikke på forskning [24, 63]. Der er lavet et udtræk fra FaktaCenteret af en patients analyseresultater for *P-Creatinin*, over en periode på 150 dage. Disse resultater er plottet på figur 4.1, for at illustrerer prøveresultaternes fordeling over perioden på 150 dage. Som det kan ses, er der få eller ingen prøveresultater tidligt i perioden. Det kan også ses, at der kan opstå klynger af data på forskellige tidspunkter i perioden. Dette er en karakteristik, der går igen for de fleste af patienterne i dette studie.

4.2 Demografi

Herunder i tabel 4.1 ses demografien for KRC-gruppen og Kontrolgruppen. Ligeledes ses også de biokemiske analysetyper, som ligger til grund for 45 af de i alt 46 features, der blev evalueret i dette projekt. Alle de features der er blevet undersøgt, kan ses i tabel 4.2. Tabel 4.1 indeholder desuden en P-værdi for de forskellige analysetyper og alderen. Er P-værdien under $0,05$ indikerer det, at der er en statistisk forskel imellem de to grupper. P-værdien kan bruges til at indikere hvilke analysetyper, der måske vil give gode resultater i klassifikatorerne.



Figur 4.1: *Dette diagram viser et eksempel på fordelingen af prøveresultater for en patient inkluderet i dette studie. 0 er starten på kontakten hvor kolorektal kræft diagnosticeres.*

Demografi	**KRC-gruppe**	**Kontrolgruppe**	**P-værdi***
Antal patienter	125	6.149	–
Antal prøveresultater pr. patient, gns. (std.)	206 (151,0)	247 (195,0)	–
Alder, gns. (std.)	71 (10,7)	62 (15,7)	0,42
Mænd/Kvinder	56/69	3.219/2.965	–

Analysetyper	**KRC-gruppe**	**Kontrolgruppe**	**P-værdi***
B-Hæmoglobin, gns. (std.)	7,0 (0,8)	7,4 (1,0)	$6,95 \cdot 10^{-8}$
P-Creatinin, gns. (std.)	102,2 (79,5)	109,6 (119,6)	0,83
P-Kalium-ion, gns. (std.)	4,0 (0,4)	4,1 (0,4)	0,20
P-Natrium-ion, gns. (std.)	138,3 (3,1)	138,5 (3,4)	0,62
B-Leukocytter, gns. (std.)	10,6 (5,4)	10,2 (7,5)	0,24
P-Albumin, gns. (std.)	33,8 (4,8)	35,1 (5,5)	0,01
P-C-reaktivt protein (CRP), gns. (std.)	59,6 (40,8)	59,9 (46,3)	0,66
P-Calcium(total), gns. (std.)	2,2 (0,1)	2,2 (0,1)	0,30
P-Calcium(total)(alb.korr.), gns. (std.)	2,4 (0,1)	2,4 (0,1)	0,90
B-Thrombocytter, gns. (std.)	352,2 (108,6)	306,6 (130,8)	$1,57 \cdot 10^{-5}$
P-Koagulationsfaktorer 2,7,10 (INR), gns. (std.)	1,6 (0,7)	1,3 (0,5)	$1,37 \cdot 10^{-7}$
P-Basisk phosphatase(BASP), gns (std.)	135,3 (128,0)	136,0 (146,9)	0,45
P-Alanintransaminase(ALAT), gns. (std.)	29,4 (26,6)	50,2 (111,4)	0,02
P-Bilirubiner, gns. (std.)	11,1 (16,7)	15,4 (30,8)	0,19
B-Erythrocytter(MCV), gns. (std.)	89,1 (7,3)	91,1 (6,9)	$2,63 \cdot 10^{-5}$

Tabel 4.1: *Demografi for KRC- og kontrolgruppen, hvor gns. er gennemsnittet af resultatværdien og std. er standardafvigelsen af resultatværdien. * Konfidensinterval 0.95*

4.3 Indledende klassifikatoranalyse

I metodeafsnit 3.2.3 beskrives de 15 forskellige analysetyper, som er brugt til at danne features. For hver analysetype er der i dette projekt tre forskellige statistiske karakteristika. Alle analyserede features ses i tabel 4.2.
På baggrund af tabel 4.2 blev der i den indledende klassifikatoranalyse, dannet klassifikatorer for alle kombinationer af 1, 2 og 3 features. Til klassifikation blev der benyttet to klassifikationsmetoder: Lineær Diskriminant Analyse (LDA) og Kvadratisk Diskriminant Analyse (QDA). Derfor vises resultater for begge klassifikationsmetoder sideløbende i det kommende afsnit.

		L. Reg.	Gns.	Std.	Normal
1.	B-Hæmoglobin	✗	✗	✗	
2.	P-Creatinin	✗	✗	✗	
3.	P-Kalium-ion	✗	✗	✗	
4.	P-Natrium-ion	✗	✗	✗	
5.	B-Leukocytter	✗	✗	✗	
6.	P-Albumin	✗	✗	✗	
7.	P-C-reaktivt protein (CRP)	✗	✗	✗	
8.	P-Calcium(total)	✗	✗	✗	
9.	P-Calcium(total)(alb.korr.)	✗	✗	✗	
10.	B-Thrombocytter	✗	✗	✗	
11.	P-Koagulationsfaktorer 2,7,10 (INR)	✗	✗	✗	
12.	P-Basisk phosphatase(BASP)	✗	✗	✗	
13.	P-Alanintransaminase(ALAT)	✗	✗	✗	
14.	P-Bilirubiner	✗	✗	✗	
15.	B-Erythrocytter(MCV)	✗	✗	✗	
16.	Alder				✗
Ialt		$15 \cdot 3 + 1 = 46$ Forskellige faktorer			

Tabel 4.2: *Tabellen viser de forskellige aggregeringer, værdier og analysetyper, som kombineres til de i alt 46 forskellige features der bruges i analysen. L.Reg. = Lineær Regression. Gns. = middelværdi. Std. = Standard Afvigelse.*

4.3.1 Klassifikatorer med 1 feature

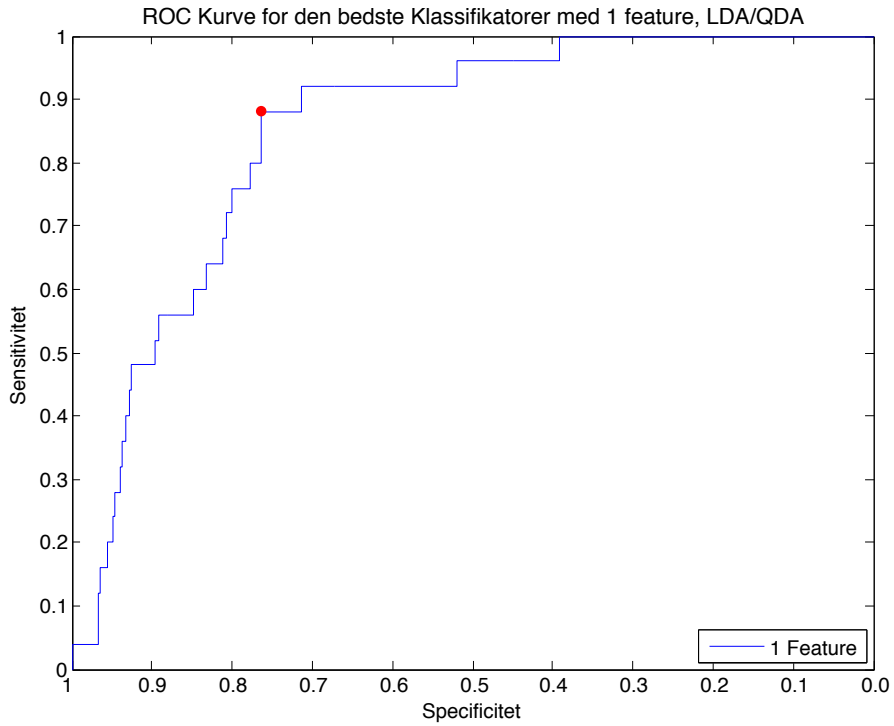
Dette afsnit præsenterer resultaterne af dataanalysen med 1 feature. Herunder er resultatet af den klassifikator, der har den højeste sensitivitet ift. specificitet præsenteret. Resultatet vises både for LDA og QDA.
I tabel 4.3 ses det sæt af klassifikatorer med den højeste sensitivitet ift. specificitet og baseret på én feature. Som det ses af tabel 4.3 er der ens værdier for både den lineære og den kvadratiske metode. Det skyldes at analyser med én feature får en beslutningsgrænse der kun er et punkt. Begge metoder finder derfor også lineær regression af *P-Albumin* som den bedste klassifikator.

Da beslutningensgrænsen kun er et punkt, er der ingen forskel i karakteristika for ROC kurverne tilhørende klassifikatorerne i tabel 4.3. Derfor vises kun en ROC kurve i 4.2. Den røde prik illustrer hvor sensitiviteten og specificteten er aflæst. I appendiks C.1 kan resultater og ROC kurver for de næst og tredjebedste klassifikatorer ses.

1. Klassifikator	**LDA**	**QDA**
Feature kombination	L.Reg. P-Albumin	L.Reg. P-Albumin
Positiv Prædiktivitet [%]	7,3	7,3
Negativ Prædiktivitet [%]	99,7	99,7
Sensitivitet [%]	88,0	88,0
Specificitet [%]	76,5	76,5
Fejlrate [%]	23,3	23,3

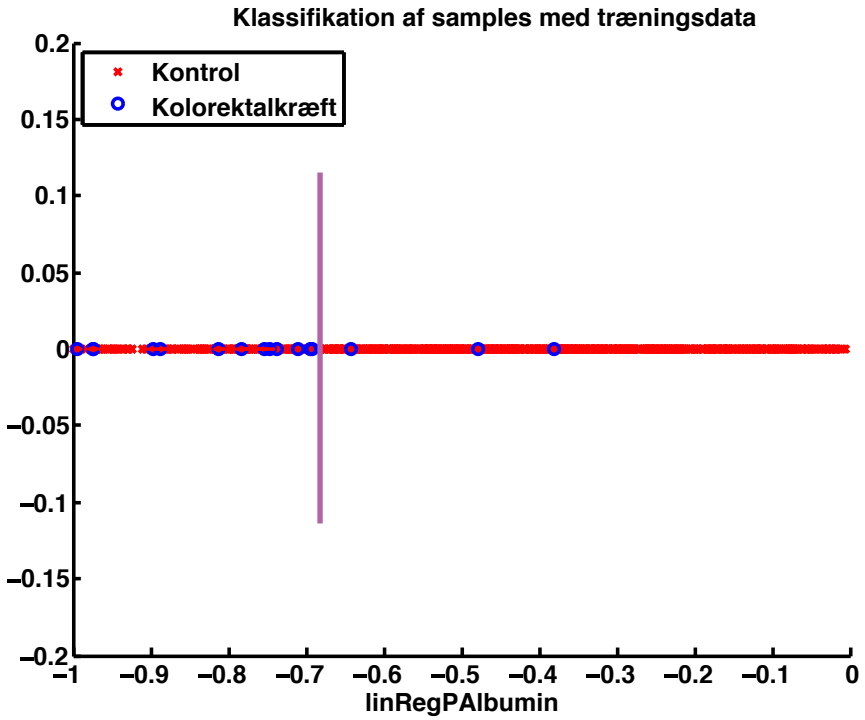
Valideringssamples	
Antal KRC	25
Antal Kontrol	1186

Tabel 4.3: *Tabellen viser de to klassifikatorer med én feature, der har den højeste sensitivitet ift. specificitet for hhv. LDA og QDA.*



Figur 4.2: *Figuren viser ROC kurven for klassifikatoren med højeste sensitivitet og specificitet baseret på én feature. Det røde punkt på linjen indikerer resultatet, der er vist i tabel 4.3*

Figur 4.3 viser beslutningspunktet for lineær regression af *P-Albumin* for både LDA og QDA. Beslutningspunktet er plottet som en linje for at fremhæve punktet, men i realiteten skærer linjen kun et sted og dermed bliver linjen til et punkt. Det ses på figuren, at der er store overlap mellem de to grupper, og at kolorektal kræft (KRC) er spredt ud over et stort område, hvorfor sensitiviteten er 88,0% og specificiteten 76,5%. Der ses dog en tendens på plottet, at patienter med KRC grupperes mod en mere negativ hældning for den lineær regression for *P-Albumin* end kontrolgruppen, hvilket indikerer et faldende albumin niveau i den undersøgte periode.



Figur 4.3: *Figuren viser beslutningspunktet for både QDA og LDA med én feature.*

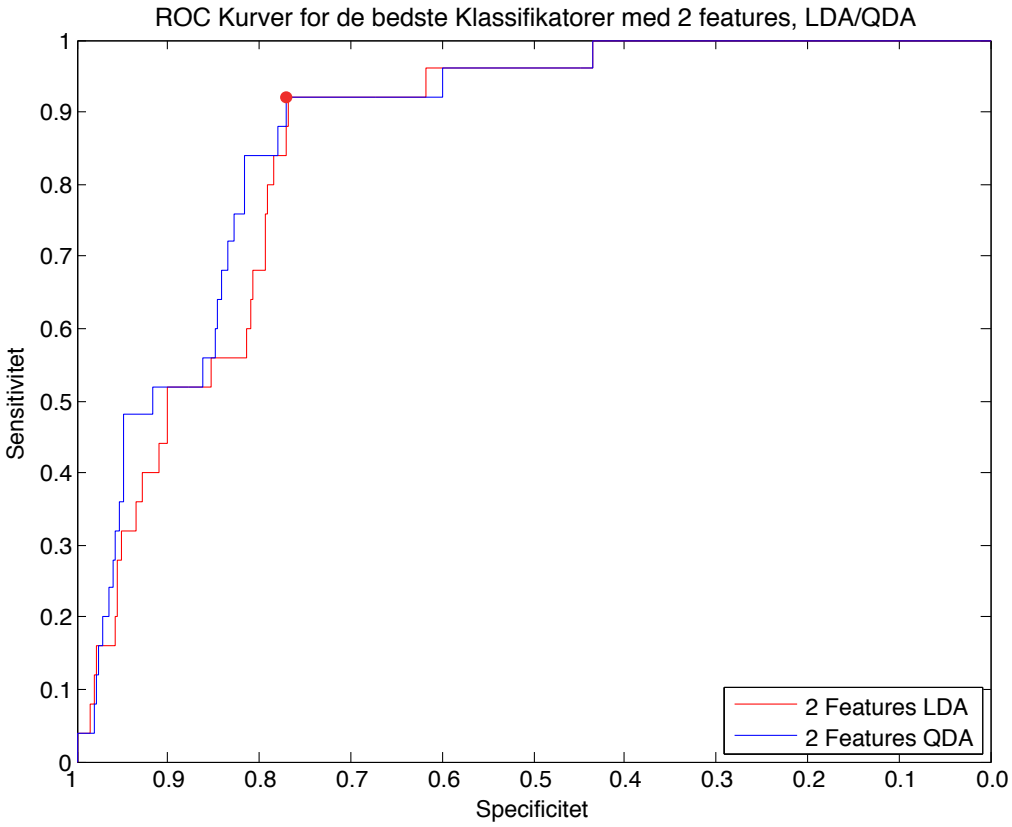
4.3.2 Klassifikatorer med 2 features

Herunder i tabel 4.4 vises de to klassifikatorer med to features, der har den højeste sensitivitet ift. specificitet, for henholdsvis den lineære og den kvadratiske analyse.
Figur 4.4 viser ROC kurver for klassifikatorerne i tabel 4.4. Den røde linje er for den lineære klassifikator, mens den blå linje er for den kvadratiske klassifikator. Resultater og ROC kurver for de næst- og tredjebedste klassifikatorer kan ses i C.1.

1. Klassifikator	**LDA**	**QDA**
Feature kombination	L.Reg. P-Albumin Gns. Erythrocytter MCV	L.Reg. P-Albumin Gns. B-Thrombocytter
Positiv Prædiktivitet [%]	7,7	7,8
Negativ Prædiktivitet [%]	99,8	99,8
Sensitivitet [%]	92,0	92,0
Specificitet [%]	76,8	77,2
Fejlrate [%]	22,9	22,5

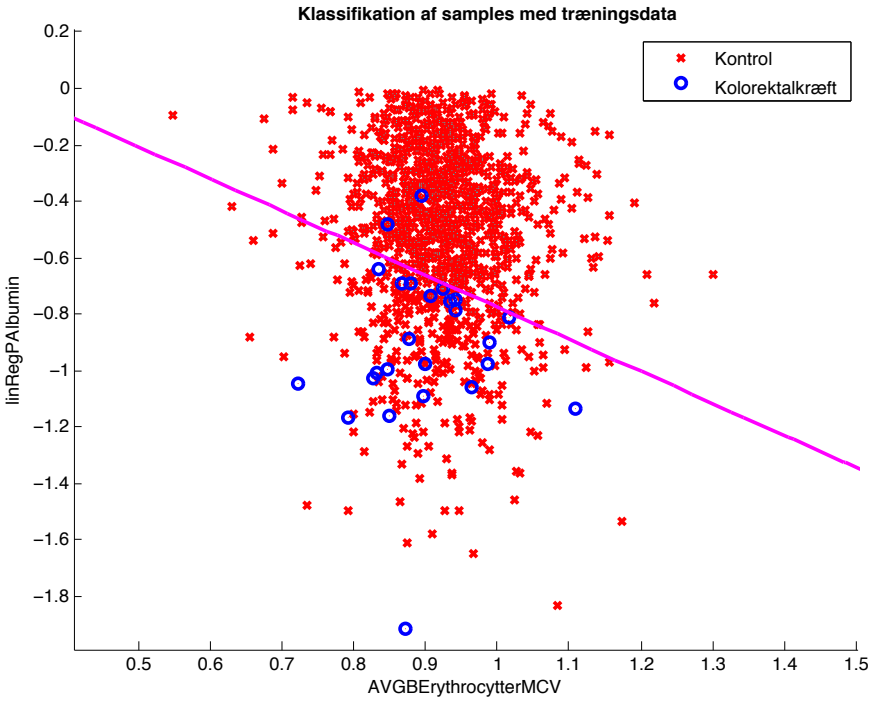
Valideringssamples	
Antal KRC	25
Antal Kontrol	1186

Tabel 4.4: *Tabellen viser de to klassifikatorer med højeste sensitivitet og specificitet baseret på to features for hhv. LDA og QDA.*

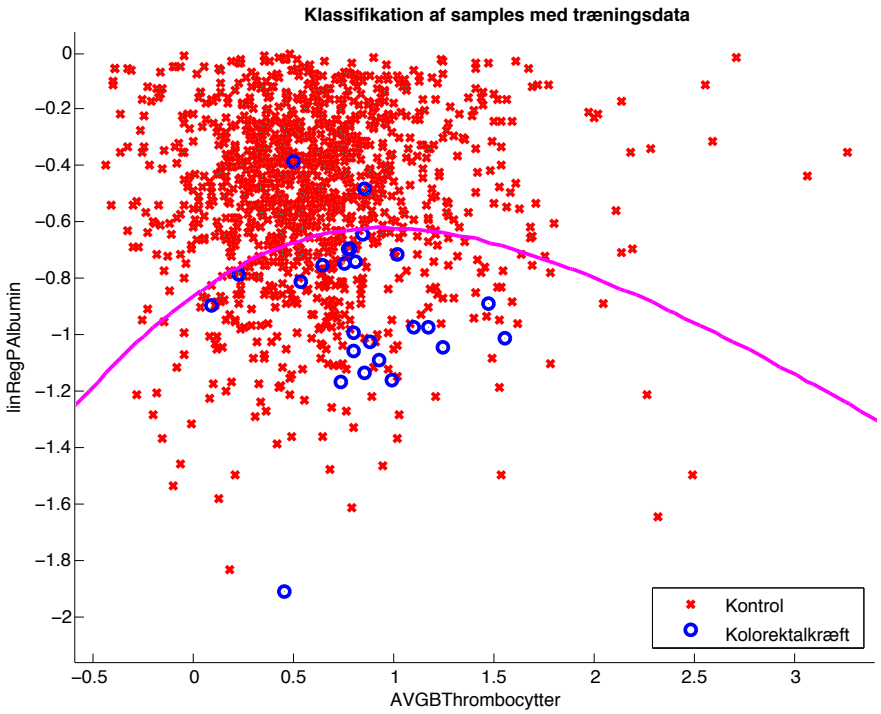


Figur 4.4: *Figuren viser de to ROC kurver for de bedste klassifikatorer med to features. Det røde punkt på kurverne indikerer resultatet der er vist i tabel 4.4*

De følgende to figurer 4.5 og 4.6 illustrer de beslutningslinjer, der er brugt til at adskille de to grupper (KRC og kontrol) fra hinanden med henholdsvis den lineære og kvadratiske analyse.
Figur 4.5 viser den linære diskriminant funktion for den lineære regression af *P-Albumin* og middelværdien (AVG) for *B-Erythrocytter(MVC)*. Lineær regression for *P-Albumin* viser en mere negativ hældning for patienter med KRC end for patienter i kontrolgruppen. For erythrocytterne ser middelværdien for begge grupper umiddelbart ens ud, men det ser ud til at erythrocytterne er lidt lavere for patienter med KRC, hvilket også afspejles i den negativ hældning for beslutningsgrænsen.
Figur 4.6 illustrer beslutningslinjen for den kvadratisk klassifikator baseret på lineær regression af *P-Albumin* og middelværdien for
textitB-Thrombocytter. Lineær regression af *P-Albumin* udviser stadig et større fald for patienter med KRC end for patienter i kontrolgruppen. Middelværdien for thrombocytterne ser ud til at gruppere patienter med KRC til højre for middelværdien af kontrolgruppen. Det indikerer at middelværdien af *B-Thrombocytter* er højere for patienter med KRC.



Figur 4.5: *Figuren viser den lineære beslutningslinje for to features.*



Figur 4.6: *Figuren viser den kvadratiske beslutningslinje for to features.*

4.3.3 Klassifikatorer med 3 features

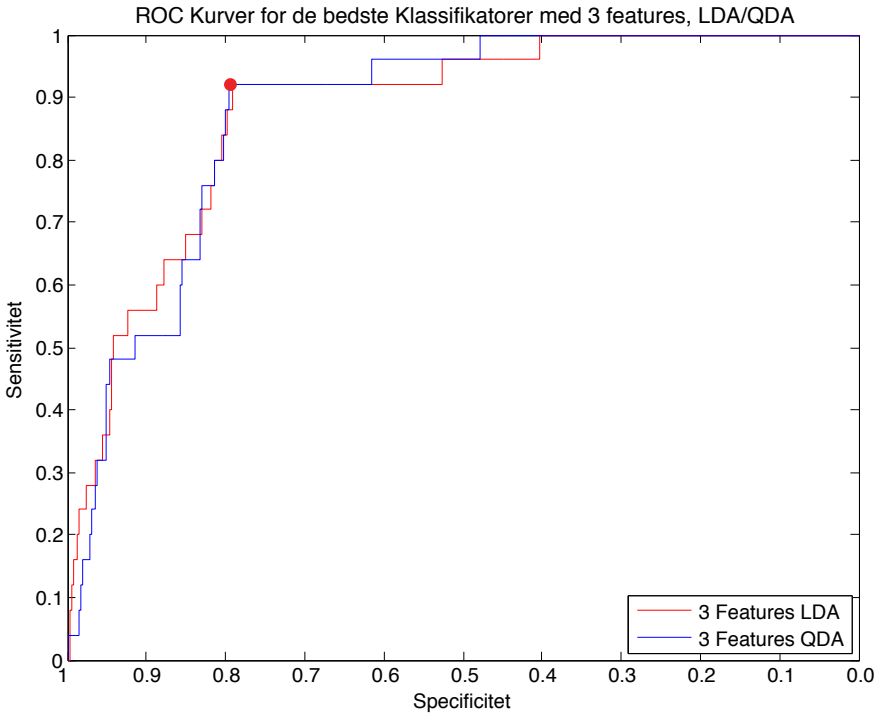
Tabel 4.5 viser de to klassifikatorer med højeste sensitivitet ift. specificitet baseret på tre features for henholdsvis den lineære og den kvadratiske analyse.
Figur 4.7 viser ROC kurver for klassifikatorerne i tabel 4.5. Den røde linje er for den lineære klassifikator, mens den blå linje er for den kvadratiske klassifikator. ROC kurver for de to resterende sæt af klassifikatorer kan ses i C.1. Det røde punkt på ROC kurverne afspejler resultaterne i tabel 4.5

1. Klassifikator	**LDA**	**QDA**
Feature kombination	Std. P-Koagulationsfaktorer L.Reg. P-Natriumion L.Reg. B-Leukocytter	Gns. Hæmoglobin Gns. B-Thrombocytter Gns. P-Koagulationsfaktorer
Positiv Prædiktivitet [%]	8,5	8,7
Negativ Prædiktivitet [%]	99,8	99,8
Sensitivitet [%]	92,0	92,0
Specificitet [%]	79,0	79,5
Fejlrate [%]	20,7	20,2

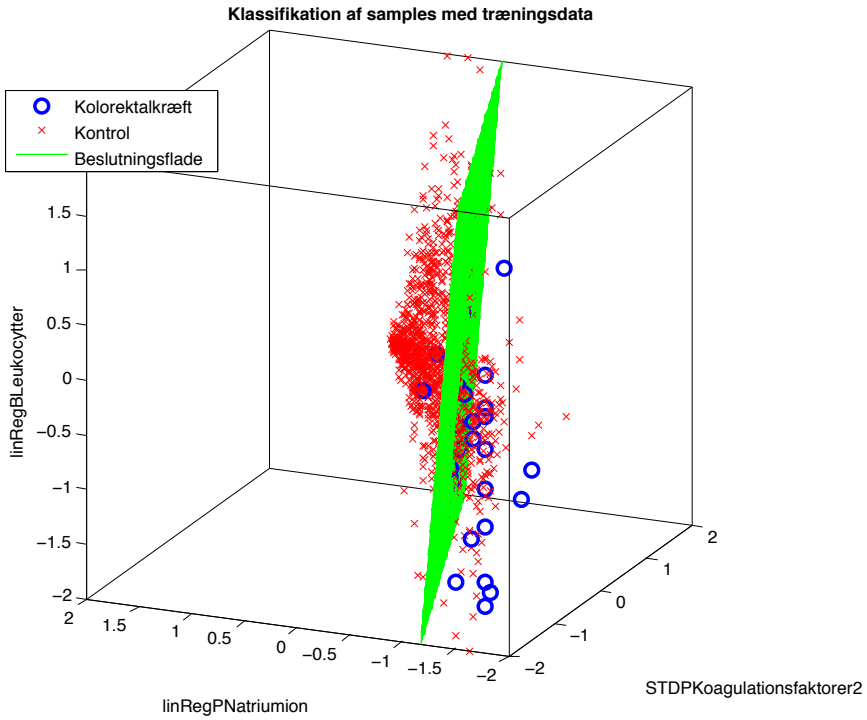
Valideringssamples	
Antal KRC	25
Antal Kontrol	1186

Tabel 4.5: *Tabellen viser de to klassifikatorer med højeste sensitivitet og specificitet baseret på tre features for hhv. LDA og QDA.*

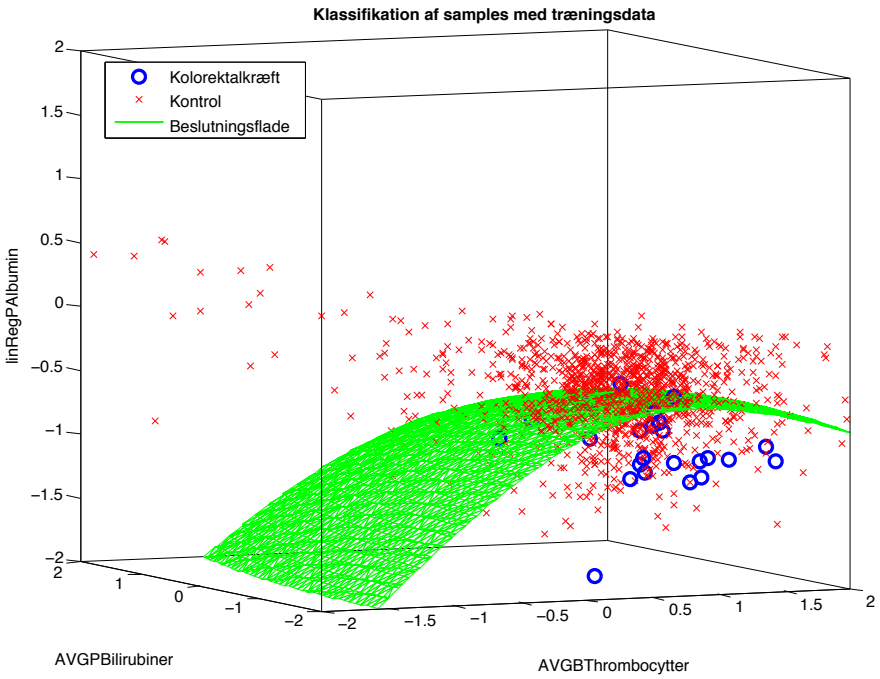
De følgende to figurer 4.8 og 4.9 illustrer de beslutningsflader, der er brugt til at adskille KRC-gruppen og kontrolgruppen fra hinanden for henholdsvis den lineære og kvadratiske analyse. Der er tre features og derfor er det et 3-dimensionelt plot. Beslutningsflader for de næst og tredjebedste klassifikatorer kan ligeledes ses i bilag C.1. Figur 4.8 viser en lineær beslutningsflade beregnet udfra den lineære regression for *B-Leukocytter*, den lineære regression for *P-Natriumion* og standardafvigelsen (STD) for *P-Koagulationsfaktorer 2,7,10 (INR)*. Den lineære regression af *B-Leukocytter* viser et fald for patienter med KRC ift. kontrolgruppen. Lineær regression af *P-Natriumion* viser også et fald for patienter med kolorektal. Standardafvigelsen for *P-Koagulationsfaktorer 2,7,10 (INR)* viser at spredningen er næsten ens for begge grupper. Figur 4.9 viser en kvadratisk beslutningsflade beregnet ud fra den lineære regression af *P-Albumin*, middelværdien for *B-Thrombocytter* og middelværdien for *P-Bilirubiner*. Lineær regression og middelværdien for *B-Thrombocytter* er allerede beskrevet i den 2 dimensionelle kvadratisk analyse, som kan ses på figur 4.6. Middelværdien for *P-bilirubiner* ser ud til at være mindre for patienter med KRC end for patienter i kontrolgruppen, selvom der er et stort overlap.



Figur 4.7: *Figuren viser de to ROC kurver for de klassifikatorer med højeste sensitivitet ift. specificitet basert på tre features. Det røde punkt på kurverne indikerer resultatet, der er vist i tabel 4.5*



Figur 4.8: *Figuren viser den bedste lineære beslutningsflade for tre features.*



Figur 4.9: *Figuren viser den bedste kvadratiske beslutningsflade for tre features.*

4.4 Afgrænset klassifikatoranalyse

Med udgangspunkt i den indledende klassifikatoranalyse er alle features fra de bedste klassifikatorer udvalgt. De udvalgte features er identificeret i tabel 4.3, 4.4, 4.5, C.1 på side 89, C.2 på side 91, C.3 på side 95. Herunder i tabel 4.6, ses de features, der er blevet udvalgt og brugt i den afgrænsede klassifikatoranalyse.
I den afgrænsede klassifikatoranalyse er der ligesom i den indledende klassifikatoranalyse, anvendt LDA / QDA til at danne beslutningsflader. I den afgrænsede klassifikatoranalyse er der dannet klassifikatorer med kombinationer af fire, fem, seks, syv, otte, ni og ti forskellige features.

		L. Reg.	Gns.	Std.	Normal
1.	B-Hæmoglobin	✗			
2.	P-Creatinin				
3.	P-Kalium-ion				
4.	P-Natrium-ion	✗			
5.	B-Leukocytter	✗			
6.	P-Albumin	✗			
7.	P-C-reaktivt protein (CRP)			✗	
8.	P-Calcium(total)		✗		
9.	P-Calcium(total)(alb.korr.)				
10.	B-Thrombocytter		✗	✗	
11.	P-Koagulationsfaktorer 2,7,10 (INR)		✗	✗	
12.	P-Basisk phosphatase(BASP)				
13.	P-Alanintransaminase(ALAT)				
14.	P-Bilirubiner		✗		
15.	B-Erythrocytter(MCV)		✗		
16.	Alder				✗

Tabel 4.6: *Tabellen viser de 13 forskellige features, der er brugt i den afgrænsede analyse. L.Reg. = Lineær Regression. Gns. = middelværdi. Std. = Standard Afvigelse.*

Tabel 4.7 og 4.8 viser en sammenligning af sensitivitet, specificitet, positiv prædiktivitet og negativ prædiktivitet for hver af de klassifikatorer med højeste sensitivitet ift. specificitet baseret på kombinationer af fire til ti features. Fra tabellerne kan det ses, at der ikke er en væsentlig forbedring af resultaterne for klassifikatorerne med mere end fire features i kombination for både LDA og QDA.
Tendens for den manglende forbedring af de enkelte klassifikatorer, baseret på mere end fire features, ses også i de to figurer 4.10 og 4.11 med ROC kurver herunder. Figurerne viser ROC kurver for de bedste klassifikatorer med kombinationer af fire til ti features ad gangen.

Klassifikatorer med 4 - 10 features for LDA

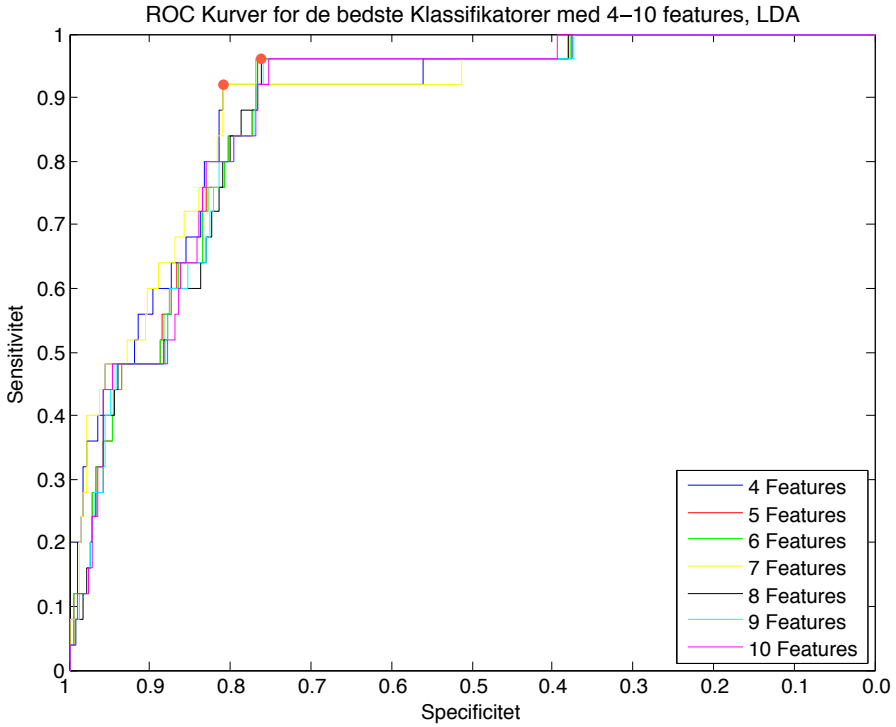
Tal er i [%]	**Sensitivitet**	**Specificitet**	**Pos. Prædiktivitet**	**Neg. Prædiktivitet**
4 Features	92,0	80,9	9,2	99,8
5 Features	96,0	76,6	8,0	99,9
6 Features	96,0	76,7	8,0	99,9
7 Features	96,0	67,0	5,8	99,9
8 Features	96,0	66,8	5,7	99,9
9 Features	96,0	76,0	7,8	99,9
10 Features	96,0	75,2	7,6	99,9

Tabel 4.7: *Tabellen viser resultater for klassifikatorerne med højeste sensitivitet ift. specificitet baseret på de udvalgte features i tabel 4.6. Resultaterne er angivet for LDA.*

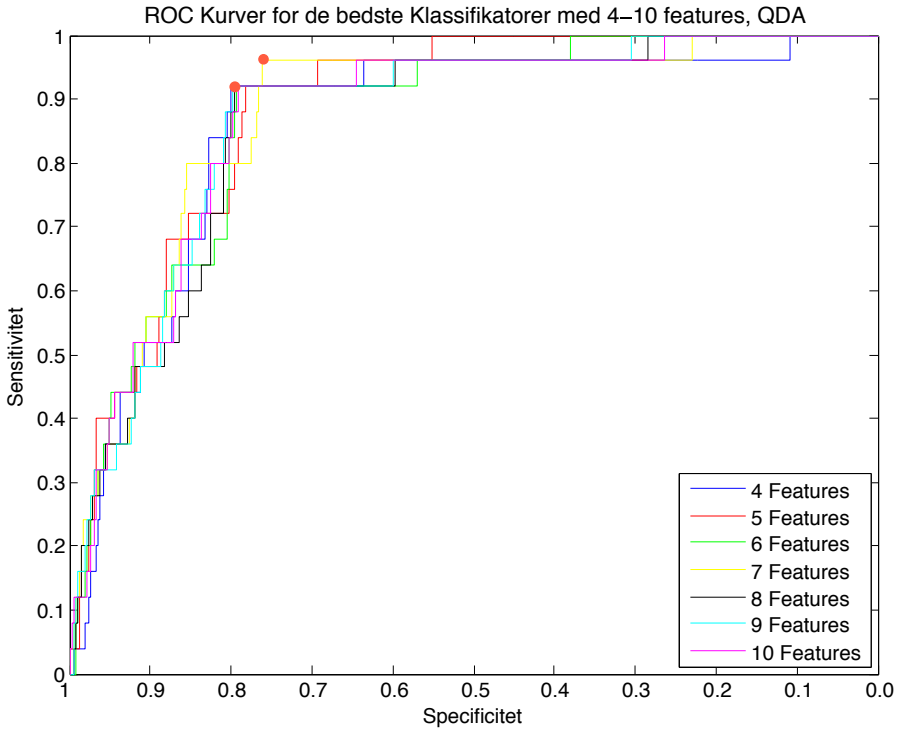
Klassifikatorer med 4 - 10 features for QDA

Tal er i [%]	**Sensitivitet**	**Specificitet**	**Pos. Prædiktivitet**	**Neg. Prædiktivitet**
4 Features	92,0	79,9	8,8	99,8
5 Features	92,0	78,3	8,2	99,8
6 Features	92,0	79,3	8,6	99,8
7 Features	96,0	76,1	7,8	99,9
8 Features	92,0	79,6	8,7	99,8
9 Features	92,0	79,8	8,8	99,8
10 Features	92,0	79,1	8,5	99,8

Tabel 4.8: *Tabellen viser resultater for klassifikatorerne med højeste sensitivitet ift. specificitet baseret på de udvalgte features i tabel 4.6. Resultaterne er angivet for QDA.*



Figur 4.10: *Figuren viser ROC kurver for de lineære klassifikatorer med 4-10 features fra tabel 4.7.*



Figur 4.11: *Figuren viser ROC kurver for de kvadratiske klassifikatorer med 4-10 features fra tabel 4.8.*

4.5 Delkonklusion

Fra den indledende og den afgrænsede klassifikatoranalyse er de to klassifikatorer med højest sensitivitet ift. specificitet for både LDA og QDA, udvalgt. De fire klassifikatorer kan ses i tabel 4.9. Fra den indledende analyse er det de to klassifikatorer baseret på tre features i kombination for både LDA og QDA. Fra den afgrænsede analyse, er det de to klassifikatorer baseret på fire features i kombination også for både LDA og QDA.
Figur 4.12 viser ROC kurver for de fire klassifikatorer fra tabel 4.10. Fra figur 4.12 og fra tabel 4.10 ses det, at klassifikatoren med den højest sensitivitet ift. specificiteten er den lineære klassifikator baseret på 4 features. Derefter følger den kvadratiske analyse med 4 features osv. Forskellen mellem klassifikatorerne med hhv. 3 og 4 features er marginal.

Klassifikatorer med 3 og 4 features og højeste sensitivitet ift. specificitet.

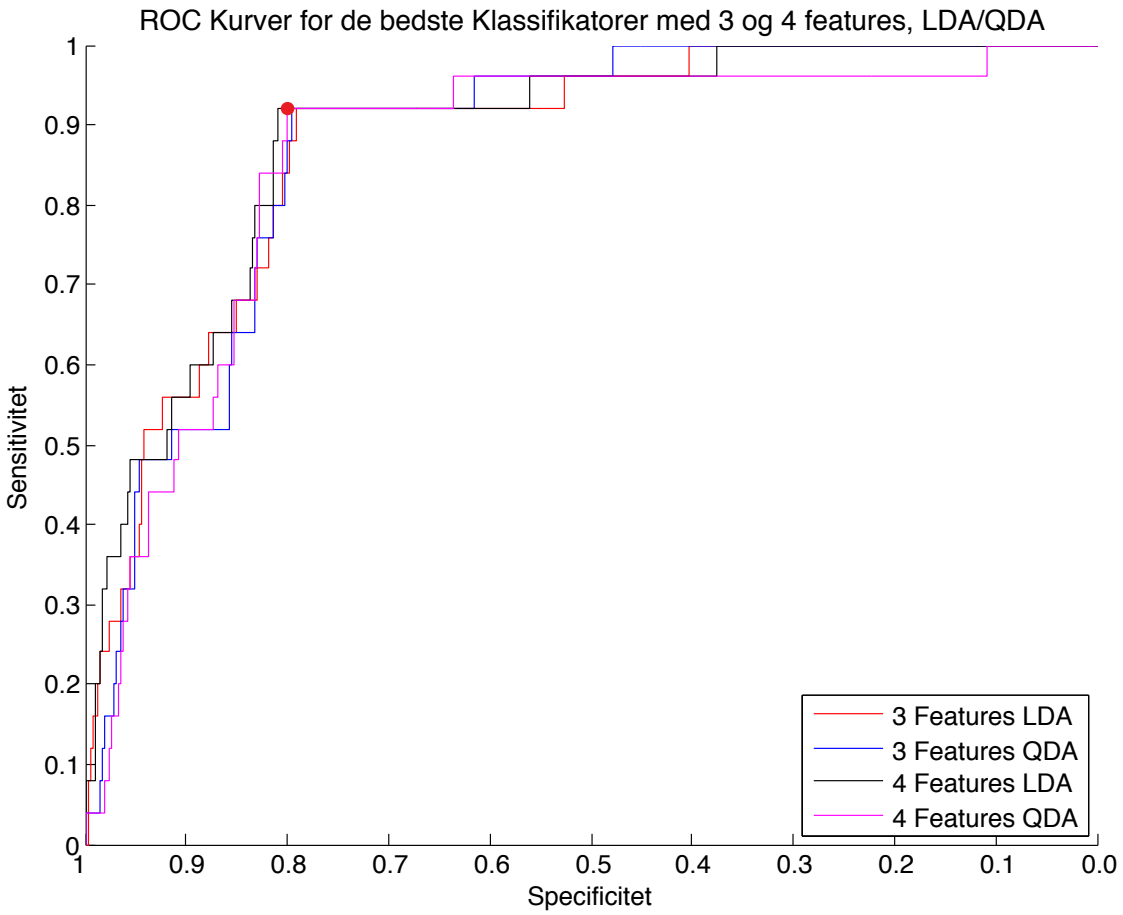
Klassifikator	**Feature kombination**
3 Features LDA	*Std. P-Koagulationsfaktorer 2,7,10 (INR)* *L.Reg. P-Natriumion* *L.Reg. B-Leukocytter*
3 Features QDA	*Gns. B-Thrombocytter* *Gns. P-Bilirubiner* *L.Reg. P-Albumin*
4 Features LDA	*Gns. B-Thrombocytter* *Std. P-Koagulationsfaktorer 2,7,10 (INR)* *L.Reg. P-Natriumion* *L.Reg. B-Leukocytter*
4 Features QDA	*Gns. B-Thrombocytter* *Gns. P-Calcium total* *Std. P-Creaktivt Protein CRP* *L.Reg. P-Albumin*

Tabel 4.9: *Klassifikatorer med højeste sensitivitet ift. specificitet baseret på 3 og 4 features for hhv. LDA og QDA.*

Klassifikatorer med 3 og 4 features og højeste sensitivitet ift. specificitet.

Tal er i [%]	**Sensitivitet**	**Specificitet**	**Pos. Prædiktivitet**	**Neg. Prædiktivitet**
3 Features LDA	92,0	79,0	8,5	99,8
3 Features QDA	92,0	79,5	8,7	99,8
4 Features LDA	92,0	80,9	9,2	99,8
4 Features QDA	92,0	79,9	8,8	99,8

Tabel 4.10: *Resultaterne af klassifikatorerne med højeste sensitivitet ift. specificiteten.*



Figur 4.12: *Figuren viser ROC kurver for de bedste klassifikatorer med højest sensitivitet ift. specifictet.*

Diskussion 5

I dette afsnit vil relevante dele af projektet blive diskuteret. Første vil der være en diskussion af udvælgelsen af de bedste klassifikatorer. Dette følges op ad diskussion af resultaterne i forhold til litteraturen med henblik på, at belyse om andre har undersøgt nogle af de sammen features og om kolorektal kræft (KRC) kan påvirke de udvalgte features. Til sidst vil der være en diskussion af begrænsningerne i dette projekt.

5.1 Udvælgelse af de bedste klassifikatorer

For at kunne finde frem til den bedste klassifikator, er det nødvendigt at definere, hvordan klassifikatorerne skal vurderes. Det er en meget svær øvelse, og det kan være svært at definere, hvad den bedste klassifikator er. Til dette er der nogle statistiske redskaber, som kan bruges. Sensitiviteten og specificiteten er redskaber, der vurderer klassifikatorens ydeevne. Sensitivitet og specificitet er værdier, der udtrykker antallet af *Type I* og *Type II* fejl. En *Type I* fejl er et falsk positivt resultat Altså hvor en patient fra kontrolgruppen bliver klassificeret som en patient tilhørende gruppen af patienter med KRC. En *Type II* fejl er et falsk negativt resultat. Altså en patient med KRC, som bliver klassificeret som værende en del af kontrolgruppen [62, 31].
Når målene til vurdering af klassifikatorernes ydeevne udvælges, er det vigtigt at have gjort to overvejelser. 1) Klassificerer man raske personer, som værende syge, kan dette lede til unødige opfølgning, men også til at den pågældende person kommer til at føle sig sygeliggjort uden grund. 2) Omvendt er det vigtigt at overveje konsekvenserne af, at erklærer en patient for rask, når patienten reelt er syg [26, 31, 62].

I dette projekt er målet at finde indikatorer, der kan detektere KRC. Detekteringen af KRC skal ses i et screenings perspektiv, og derfor ønskes en høj sensitivitet. Men der er også behov for en høj specificitet, ellers vil for mange patienter skulle have opfølgende undersøgelser. Grunden til at en høj sensitivitet ønskes er, at vi ønsker, at detektere alle de personer, der har KRC. Havde man eksempelvis en lav sensitivitet på 60%, så ville 40% af patienterne med KRC ikke blive detekteret. Øges sensitiviteten, er det vigtigt at tage højde for, at specificiteten vil være lavere. Dermed klassificeres raske personer som syge [31].
Sensitiviteten og specificiteten skal optimeres med en overvægt til sensitiviteten. Specificiteten vægtes lavere ift. sensitiviteten, men spørgsmålet er hvad er den laveste acceptable specificitet? Det er svært at lave denne optimering, så derfor bruges ROC kurverne, fordi de ofte afslører et naturligt punkt for optimeringen. ROC kurven blev derfor brugt til at aflæse sensitivitet og specificitet manuelt. Udvælgelsen af den bedste sensitivitet ift. specificiteten gøres altså ud fra

nogle faktorer, som kan være svære at kvantificere og som er kontekst afhængige. Derfor vil udvælgelsen afspejle forskernes fokus.

Kompleksiteten af klassifikatorerne er også en vigtig faktor at vurdere når den bedste klassifikator skal vælges. Overtræningsfænomenet, beskrevet tidligere i metoden, spiller ind i denne beslutning. Det kan være at en højere dimensionalitet giver en marginalt bedre klassifikator, men kompleksiteten gør måske at ydeevnen udjævnes desto flere samples, der klassificeres. Derfor er den generelle regel jvf. *Ockhams Razor*, ofte at vælge den simpleste klassifikator af to, der yder relativt ens [58].

Den bedste klassifikator udvælges derfor i dette projekt efter en optimeret sensitivitet, specificitet og kompleksitet. Sensitivitet og simplicitet vægtes højt. Pointen er at vægtningen af sensitivitet, specificitet og kompleksitet er en subjektiv vurdering, der kræver erfaring og forståelse af konsekvenserne.

5.2 Resultater

I dette projekt blev der gennemført en indledende og en afgrænset analyse, hvor der i den indledende analyse blev beregnet klassifikatorer for alle mulige kombinationer af én, to og tre features ad gangen for alle 46 features. Beregning af klassifikatorer for kombinationer af fire eller flere features var tidskrævende. Derfor var det kun features fra de tre bedste klassifikatorer fra den indledende analyse, som blev anvendt til beregning af klassifikatorer for alle mulige kombinationer af fire og op til ti features. Denne afgrænsning i antallet af features betød at det var væsentligt hurtigere at undersøge hvilke kombinationer, der gav de bedste resultater.
Svagheden i denne tilgang er, at det kun er de bedste features fra den første del af analysen, som bruges i den videre analyse. Nogle af de øvrige features kunne risikere at udviser bedre diskriminativ karakter ved en højere dimensionalitet.

Den indledende analyse viste, at den tre dimensionelle LDA og QDA havde højeste sensitivitet ift. specificitet. Den afgrænsede analyse viste jvf. tabel 4.7 og 4.8, at klassifikatorer med mere end fire features ikke giver bedre resultater end dem med tre og fire features. Denne manglende forbedring af resultaterne ved dimensionaliter over fire indikerer, at de resterende features ikke bidrager med yderligere væsentlig information, der kan bruges til at adskille patienter med KRC fra patienterne i kontrolgruppen.

Ud fra tabel 4.10 på side 54 og figur 4.12 på side 55 ses det, at der kun er en marginal forbedring af sensitivitet og specificitet ved at øge kompleksiteten af klassifikatorerne fra tre til fire features. Det kan derfor diskuteres, om det kan betale sig at øge kompleksiteten fra tre til fire features for at opnå den marginale forbedring. Med hensyn til et nyt dataudtræk fra datavarehuset, så vil færre analysetyper kunne give flere samples. Derudover er forskellen på de tre og fire dimensionelle analyser så lille, at det kan være et tilfælde. Derfor vælges de tre-dimensionelle klassifikatorer, som de to bedste.

Den tre dimensionelle lineære klassifikator udnytter standardafvigelsen af *P-Koagulationsfaktor 2,7,10 (INR)*, lineær regression af *P-Natriumion* og *B-Leukocytter*. Den opnår en sensitivitet på 92,0% og specificitet på 79,0, jvf. tabel 4.10 på side 54.
P-Koagulationsfaktor 2,7,10 (INR) kan muligvis bruges til at indikere KRC, fordi de flest tumorer og polypper bløder sporadisk [10]. Da blødninger kan give et øget forbrug af *P-Koagulationsfaktor*

2,7,10 (INR), kan dette forklare at det netop er spredningen, som klassifikatoren udnytter [64]. Beskrivelsen af figur 4.8 afslører dog, at der er meget lille forskel på de to grupper, hvilket antyder at *P-Koagulationsfaktor 2,7,10 (INR)* ikke er en diskriminativ feature.
Mht. til *P-Natriumion*, så er indtagelse af store mængder salt blevet kædet sammen med risicien for at udvikle KRC [10]. Resultaterne 4.1 på side 42 viser, at der ikke er forskel på gennemsnittet af de to grupper for *P-Natriumion*, men figur 4.8 på side 49 viser et svagt fald i *P-Natriumion* ift. til kontrolgruppen. Litteraturen giver dog ingen direkte forklaring på dette fænomen.
En interessant observation ved denne klassifikator er, at figur 4.8 viser, at KRC-gruppen har faldende værdier for *B-Leukocytter* sammenlignet med kontrolgruppen. Dette kan måske forklares af at tumorer resulterer i blødning, og at der derfor sker et tab af leukocytter, som resulterer i de faldende værdier. Dette er dog ikke noget, der er beskrevet i litteraturen. Litteraturen beskriver til gengæld, at et forhøjet niveau af *B-leukocytter* kan kædes sammen med gastrointestinal kræft[65]. Men niveauet behøver ikke nødvendigvis være lavt, selvom resultaterne i dette projekter viser, at det er faldende. *B-Leukocytter* er en del af kroppens immunforsvar, og det reagerer også på kræft. *B-Leukocytter* vil ligesom ovennævnte analyser ikke kunne bruges alene, da infektioner og sygdom i kroppen udover kræft også vil påvirke værdierne. Eksempelvis kan kraftig aktivitet også øge niveauet af *B-leukocytter* [65].

Den tre dimensionelle kvadratiske klassifikator udnytter lineær regression af *P-Albumin* samt gennemsnittet af *B-Thrombocytter* og *P-Bilirubiner*. Den opnår en sensitivitet på 92,0% og en specificitet på 79,5%, jvf. tabel 4.10 på side 54.
En af grundene til at *B-Thrombocytter* muligvis kan bruges som indikator, er at de fleste tumorer og polypper bløder [10], hvorfor der er behov for at lukke disse sår. *B-Thrombocytter* eller blodplader er ansvarlige for de primære processer i kroppen, der stopper blødning [66]. Fælles for både *B-Thrombocytter* og *P-Koagulationsfaktor 2,7,10 (INR)* er, at de kan påvirkes af andet end kræft, hvorfor de enkeltvis ikke kan anvendes som indikatorer.
Et studie af patienter med KRC fandt, at *P-Albumin* var signifikant lavere hos patienter med KRC [38]. Det stemmer overens med resultaterne på figur 4.5 på side 47, der beskrives som faldende. Artiklen giver ingen forklaring på det lave albumin niveau. Albumin produceres i leveren og kan udskilles igennem tarmen, så det kan måske forklare forskellen mellem patienter med KRC og patienter uden. En anden grund til det lave niveau, kan være at tumoren har metastatiseret til leveren. Yderligere kan albumin indikere problemer med nyrerne, hvilke kan betyde, at kræften evt. er metastatiseret dertil.
Mht. *P-Bilirubiner* har et studie vist, at lave bilirubin-værdier korrellerer med KRC. Studiet giver ingen forklaring, det konkluderer at denne korrelation skal undersøges nærmere [67].

De bedste klassifikatorer havde, som nævnt i resultatafsnittet, en sensitivitet på 92%, hvilket giver en sensitivitet, der ligger på højde med iFOBT og gFOBT. Det skal dog nævnes at de rapporterede sensitiviteter for FOBT varierer. Hvilket betyder at klassifikatoren potentielt kan have en bedre sensitivitet. Specificiteten i dette projekt er på ca. 79%, og er dermed en del lavere end den specificitet, der er afdækket for både gFOBT og iFOBT jvf. afsnit 2.1. Dvs. at resultaterne i dette projekt sammenlignet med de screeningsmetoder, der er analyseret i 2.1, viser at detektering af KRC ved hjælp af de biokemiske markører, der er fundet igennem datavarehuset, vil have et højere antal af falske positive end de nuværende metoder. Et højt antal falske positive vil betyde, at flere personer skal undersøges yderligere for KRC. Dette kan betyde, at der skal udføres flere koloskopier. Det kan være uhensigtsmæssigt både i et økonomisk- og i et patient-

perspektiv. På den anden siden har klassifikatorerne i dette projekt potentiale til at detektere en større andel af de patienter, der har KRC, og øger dermed potentialet for tidlig indsats hos flere patienter.

5.3 Begrænsninger

Der er undervejs i problemanalysen, metoden, resultaterne og diskussionen berørt en række problemstillinger og begrænsninger i dette projekt. Disse vil bliver diskuteret i detaljer i dette afsnit.

I problemanalysen adresseres en række problemstillinger ift. til at omdanne eksisterende data til en form, der kan bruges til sekundære formål. I forbindelse med genbrugen af disse data til forskning er der yderligere nogle begrænsninger. Først og fremmest er data i datavarehuset retrospektive, dvs. at der er ingen eller meget lidt kontrol over datakilden [46]. Det betyder, at alle datatyper, som man vil bruge, ikke altid er tilstede. F.eks. ses det i datavarehuset, at meget få patienter med kolorektal kræft (KRC) har fået udført tre prøver af den samme type i året før diagnosen. En af svaghederne ved dette studiedesign er, at hver analysetype skal have minimum tre prøveresultater, før de kan omdannes til en lineær regression. Det er en svaghed, fordi det nedsætter antallet af patienter, som kan inkluderes i dataanalysen, og det nedsætter antallet af analysetyper, som kan inkluderes. Begge påvirker styrken af metoden negativt.
Den tilgængelige datamængde pr. patient påvirkes yderligere, fordi studiet er designet til at analysere data, der ligger før patientens diagnose. Det er et problem, fordi der ofte er færre data i perioden før diagnosen end i perioden lige efter diagnosen, hvor patienten er i kontakt med sundhedsvæsnet. Det er nødvendigt at begrænse perioden til året før diagnosen i et detekteringsperspektiv, fordi det er mindre relevant at bruge data efter en diagnose til at detektere diagnosen. Håbet var, at der var data til rådighed, fordi patienterne var igennem et udredningsforløb før den endelige diagnose blev stillet. Kun 125 ud af 9.915 patienter med KRC havde nok data, mens 6.149 ud af 391.679 patienter uden kræft havde nok data. De 125 patienter er ifølge et opslag i tabellen: *Scientific Table by K.Diem and C. Leitner Ciba-Geigy Ltd. 1975* stadig et godt statistisk grundlag, hvor der med 98% sikkerhed kan generaliseres om 90 procent af befolkning med KRC [68]. Ved opslag i samme tabel ses, at 6.149 patienter er nok til, at der med 98% sikkerhed kan generaliseres om 99,9% af befolkningen. Selv om opslaget fortæller, at antallet af patienter i dette projekt er et godt statistisk grundlag, så tabes der stadig en meget stor procentdel af patienterne i datavarehuset, fordi de ikke har nok data.
Det kan måske konkluderes at udredningen af KRC ift. biokemiske analyser ikke er særlig konsistent, eller at biokemiske analyser ikke er en del af det normale udredningsforløb for KRC. Det antyder at identificering af biomarkører, der kan detektere sygdomme, er uhensigtsmæssigt i et datavarehus, fordi der ikke er nok data. Dette projekt demonstrerer dog, at det kan lade sig gøre, men det lave antal af samples svækker metoden. Datavarehuset som datakilde har potentiale til at facilitere indledende dataanalyser, der kan guide senere forskning.
Fordelingen af prøveresultater på et år er illustreret på figur 4.1. Når der arbejdes med sekundære data, er det ikke muligt at kontrollere frekvensen af disse prøveresultater. Det ses på figuren, hvordan prøverne kan være fordelt tilfældigt udover perioden. Det kan være problematisk, hvis de tre prøveresultater er meget tæt på diagnosetidspunktet eller omvendt,

fordi så vil man ikke kunne detektere en hældning (ændring) over tid. Når prøverne ligger meget tæt på hinanden er lineær regression også meget sensitiv overfor varians i prøverne. Så tre prøver, der ligger meget tæt, kan også give en meget stor hældning ved en lineær regression, selvom hældningen burde være lille. Det er en faktor, som vil påvirke resultatet.

Kontrolgruppen i dette studie kan påvirke resultaterne, fordi den består af syge patienter. En del af sygdommene, som patienterne har, vil måske påvirke nogle af de samme analyser, som KRC gør. Det betyder, at specificiteten bliver påvirket negativt. I et fremtidigt studie ville det være interessant af undersøge raske patienter ift. patienter med KRC, for at se om specificiteten øges.
Når en del af de syge patienter overlapper med KRC, bliver man nødt til at overveje det fænomen, der kaldes *confounding factors* [3]. Det er et fænomen, som skal overvejes, når man finder ukendte sammenhænge i data vha. data mining og andre statistiske analyser. Data mining algoritmerne søger at adskille de to grupper så godt som muligt. Algoritmerne kan ikke diskriminere mellem data, der kan adskille KRC fra andre sygdomme, og data som kan adskille patienterne med KRC fra de andre patienter i dette studie. *Confounding factors* betyder i bund og grund, at algoritmen har fundet sammenhænge i data, der ikke har noget med KRC at gøre. F.eks. kan det være at alle de patienter, som vores algoritme klassificerer til at have KRC, har noget andet tilfældes, eksempelvis en række sygdomme, som påvirker de udvalgte analysetyper ens. Når datavarehuset ikke giver adgang til mere detaljeret data, kan det være svært at afdække *confunding factors*. I et fremtidigt perspektiv vil det være interessant at gentage denne dataanalyse med et datavarehus, som indeholder data om medicinering og data fra EPJ. Både for at identificere eventuelle *confounding factors* og/eller finde andre faktorer, der kan klassificere KRC.

Datakilden til et datavarehus skal være af god kvalitet, ellers påvirker dette resultaterne. I denne forbindelse tænkes på fejlregistreringer og manglende registrering af diagnoser for patienterne. Data fra primære kilder kan være både ikke komplette og inkonsistente. Kompletheden af data har været diskuteres i forbindelse med antallet af tilgængelige prøver. Men kompletheden kan også være registreringen af diagnoser i PAS. Hvis f.eks. en gruppe patienter ikke er blevet registret med diagnosekoden for KRC, kan det påvirke træningen af algoritmerne. En anden situation kan også opstå, hvor f.eks. KRC er blevet registreret som aktionsdiagnosen, selvom KRC var henvisningsdiagnosen (Patienten fik en anden diagnose end KRC, men var blevet henvist til en udredning for KRC). Det vil også påvirke træningen af algoritmerne. Denne problemstilling eksemplificeres ved en valideringsundersøgelse af Lands Patient Registeret (LPR). PAS bruges til at rapportere til LPR, så derfor kan validiteten af LPR overføres til PAS. Tal fra *Danish Colorectal Cancer Group's* kliniske database viste, at der var registreret 19.545 tilfælde af KRC fra 2001-2006, mens der i LPR var registreret 22.743 tilfælde af KRC i sammen periode. Det giver en afvigelse på ca. 14% [49], og det sætter spørgsmålstegn ved at bruge data fra PAS til sekundære formål. Er fejlene i den kliniske database eller i PAS?

De ovenstående begrænsning kommer pga. det retrospektive datasæt. Det er meget svært at drage endelige konklusioner fra dataanalyser pga. af mangel på kontrol over datakilder. Derimod kan retrospektive studier bruges til at danne foreløbige konklusioner, som kan lede til hypoteser. Hypoteserne kan derefter blive omdrejningspunktet for designet af stærkere prospektive studier, som kan bruges til at drage konklusioner om hypoteserne[55, s. 1260-1].

Konklusion 6

I dette kapitel gives en konklusion på projektets problemformulering. Konklusionen tager udgangspunkt i metodevalg, de udvalgte resultater, samt diskussionen af resultaterne og metodens begrænsningerne. Problemformuleringen for dette projekt var:

Hvordan er det muligt, at identificere kemiske biomarkører til detektering af kolorektal kræft, givet et datavarehus med patientadministrative data og laboratoriedata?

Mønstergenkendelses paradigmet blev anvendt i dette projekt, da formålet var at identificere nye sammenhænge i data, der kunne blive til kemiske biomarkører for kolorektal kræft (KRC). Region Nordjyllands (RN) datavarehus, FaktaCenter, blev anvendt som datakilde, hvilket medførte sekundær anvendelse af data. RN's datavarehus indeholder patientadministrative data og laboratoriedata. I forbindelse med at skabe ny viden fra sekundære data introduceres integrerings- og kontekstafhængighedsproblemstillinger, som er velkendte indenfor klinisk informatik. Klassifikationssystemer og datavarehusets underliggende model giver mulighed for at adressere begge problemstillinger, men det er af afgørende at overveje betydningen af konteksten og integreringen i metoden til mønstergenkendelse.
Laboratorierne i RN har internt nogle normalværdier, som skal følges, disse normalværdier er ikke nødvendigvis ens fra laboratorium til laboratorium. Derfor skal resultatet af en prøve fortolkes i konteksten af normalværdien for det pågældende laboratorium. For at overkomme denne problemstilling blev prøveresultaterne normaliseret.
Det betød, at der kunne udtrækkes 1 års data, fra før KRC blev diagnosticeret hos den enkelte patient, på de 15 hyppigst anvendte biokemiske analyser på tværs af alle laboratorier i RN. Til kontrolgruppen blev der også udtrukket 1 års data fra patienter uden kræft. Dette illustrer Integreringsproblemstillingen. Integrering af data er afhængigt af formålet med dataanalysen. I dette projekt var der fokus på detektering af KRC, hvilket betød, at det kun var prøveresultater i en bestemt periode, der skulle integreres med diagnosekoden. Data i datavarehuset er derfor ikke integrerede, selvom data er implementeret i den sammen underliggende model. Datavarehuset faciliterer derimod integrering, som derefter skal udføres, så den er meningsfuld ift. det enkelte sekundære formål.

For hver analysetype blev der udregnet tre forskellige statistiske mål: lineær regression, middelværdi og standardafvigelse. Det blev gjort for at identificere tendenser i data, som et enkelte datapunkt ikke kan beskrive. Igen skal de enkelte datapunkter i den udvalgte periode, ses i kontekst med de øvrige datapunkter i perioden. I alt dannede disse statistiske karakteristika 45

forskellige features. I alt blev 46 features brugt i analysen, hvor den sidste feature var patientens alder. Der blev gennemført analyser med kombinationer af én til ti forskellige features ad gangen. Mønstergenkendelsen tilbyder en lang række af metoder afhængigt af data. I dette projekt ledes der efter nye sammenhænge, så hverken *class conditional densitity* eller dens form var kendt. Derudover er data inddelt i to kendte klasser. På baggrund af disse informationer blev en lineær og en kvadratisk diskriminant analyse udført.
Der blev udført en indledende analyse med alle 46 features. Derefter blev en afgrænset analyse udført, hvor kun de 13 bedste features fra den indledende analyse blev brugt til at undersøge dimensionaliteter højere end tre.

Resultaterne af dette projekt viste, at den bedste klassifikator var en tre-dimensionel kvadratisk diskriminant analyse baseret på *Gns. B-Thrombocytter*, *Gns. P-Bilirubiner*, *P-Albumin*. Klassifikatoren giver en sensitivitet på 92 % og en specificitet på 79,5%. Sammenlignes resultaterne for den bedste klassifikator i dette projekt med de rapporterede resultater for FOBT, så er klassifikatorens specificitet dårligere, mens sensitiviteten ligger på højde med de rapporterede sensitiviteter for FOBT. Der bliver dog rapporteret meget varierende sensitiviteter for FOBT, så klassifikatoren har potentielt en bedre sensitivitet.

I dette projekt er den bedste klassifikator udvalgt efter den lavest mulige kompleksitet samt den højest mulige sensitivitet ift. specificitet. Sensitiviteten vægtes højest ift. specificteten, fordi det betyder at færrest mulige tilfælde af KRC overses. Det betyder dog også, at en større del af de raske patienter skal igennem opfølgende undersøgelser. Lav kompleksitet vægtes også højt, fordi en høj kompleksitet kan medføre fejl, og den kræver flere data.

Der er i projektet blevet beskrevet en række begrænsninger, samt problematikker man skal være opmærksom på ved anvendelse af eksisterende data i datavarehuse til sekundære formål. Både kontekstafhængighed og integrering er blevet adresseret. Sekundær anvendelse af data betyder, at datasætter er retrospektivt. Dvs. at man som forsker ikke har kontrol over datakilden. Det kommer til udtryk i dette projekt ved det lave antal af samples, som skyldes fordelingen og kompletheden af prøveresultaterne. Det er ikke alle patienter, der har fået taget alle 15 analyser tre gangen på ét år. Prøveresultaterne fra de patienter, som opfylder kravet, er ikke fordelt ens. Dvs. at alle prøveresultaterne fra en analysetype kan være grupperet tæt, og dermed kan beregningen af de statistiske karakteristika nemmere blive påvirket af måleusikkerheder.
Opslag i *Scientific Table by K.Diem and C. Leitner Ciba-Geigy Ltd. 1975* viser dog, at de 125 patienter med KRC og de 6.149 stadig er et godt statistisk grundlag.

Dette projekt har demonstreret at Region Nordjyllands datavarehus, som indeholder data fra de to kildesystemer PAS og LABKAII, kan bruges til at integrere og håndtere konteksten af data. Dermed kan data anvendes til det sekundære formål, at identificere kemiske biomarkører for KRC vha. mønstergenkendelse.
Det konkluderes derfor, at mønstergenkendelse kan anvendes på det retrospektive datasæt i datavarehuset til at identificere potentielle kemiske biomarkører for KRC. Trods begrænsningerne ser det ud til, at klassifikatoren kan adskille patienter med KRC fra patienter uden kræft. Den retrospektive karakter af projektet, gør at konklusionen ikke kan drages endeligt. Det anbefales derfor, at yderligere data om f.eks. medicinering og mere detaljerede data om diagnoserne inkluderes i datavarehuset for at afdække eventuelle *confounding factors*. Det anbefales

også, at der laves et nyt udtræk med færre analysetyper fra datavareshuset for at øge antallet af patienter, der kan inkluderes. Yderligere anbefales det, at konklusionen i dette projekt bruges, som omdrejningspunkt i et prospektivt studie. Datavarehuset er dermed en datakilde, som har potentiale til at facilitere indledende dataanalyser, der kan guide senere forskning.

7 Perspektivering

Der er to interessante perspektiver for datavarehuset, som vil blive behandlet i dette kapitel. Først perspektiveres over brugen af datavarehuset til detektering af kolorektal kræft (KRC). Dernæst perspektiveres over datavarehusets fremtidige potentiale ift. andre kliniske problemstillinger.

Resultaterne af dette studie viste, at det er muligt at anvende de data, der er i datavarehuset, til at identificere kemiske biomarkører, som kan bruges til detektering af patienter, der har KRC. Men resultaterne viste også at rigtig mange patienter fra kontrolgruppen blev kategoriseret som syge, og samtidige var det ikke muligt at finde alle patienter med KRC. Derudover konkluderes også at datavarehuset ikke umiddelbart kan bruges til screening, fordi der er for få patienter, der har nok data til at blive undersøgt. Skal vi alligevel sammenligne datavarehuset med de eksisterende metoder, så udvides tabel 2.1 på side 9 til følgende tabel 7.1 med fordele og ulemper for datavarehuset. Datavarehuset skal i denne sammenhæng levere data, som skal bruges af den bedste klassifikator. I praksis når klassifikatoren er trænet, vil det desuden være hurtigt og nemt at undersøge data fra patienter, der kan have KRC. Andre fordele er, at data allerede eksisterer i datavarehuset pga. af kontakter til sundhedsvæsnet. Det er minimalt invasivt at få taget blodprøver, men for at datavarehuset skal fungere, så skal der tages minimum tre blodprøver om året pr. patient, da klassifikatoren kræver store mængder data. Derudover er der endnu ikke evidens for klassifikatoren. Disse fordele og ulemper og manglende evidens for metoden, gør at klassifikatoren på baggrund af eksisterende data, ikke er en mulig screeningmetode. Derimod hvis metoden blev implementeret, så der blev leveret blodprøver minimum tre gange om året pr. borger, antyder resultaterne at klassifikatoren kan bruges til screening.

Resultaterne i dette projekt viste, at data mining i Region Nordjyllands datavarehus kan bruges til at identificere biomarkører for KRC. Datavarehuset indeholder på nuværende tidspunkt kun data fra to datakilder (PAS og LABKAII). Hvis datavarehuset f.eks. udvides med data om medicinering, data fra dødsårsagsregistret og/eller DCCG's kliniske database, så er der et endnu større datagrundlag, som kan bruges til at identificere nye sammenhænge og evt. eliminerer *confounding factors* og dermed øge validiteten af klassifikatoren.
Et helt konkret forslag i forhold til at inkludere flere datakilder er det kommende screeningprogram for KRC, hvor iFOBT er blevet anbefalet. Det vil være interessant at inkludere prøverne i datavarehuset, for evt. at kunne øge specificiteten i metoden i dette projekt og sensitiviteten for FOBT. Sagt med andre ord, måske indeholder datavarehuset faktorer, som sammen med iFOBT kan give bedre resultater.

Metode	Fordele	Ulemper
gFOBT	+ Noninvasiv + Nem at udføre + Manuel aflæsning + Lav stk. pris + Høj specificitet	- Diæt Restriktioner - Medicin Restriktioner - Manuel aflæsning - 3 afføringsprøver - varierende sensitivitet
iFOBT	+ Noninvasiv + Maskinaflæsning + Ingen Diæt Restriktioner + Ingen Medicin Restriktioner + Nem at udføre + Lav stk. pris	- Skal sendes til laboratorium - Maskinaflæsning - Varierende sensitivitet
Fæces DNA test	+ Noninvasiv + Ingen Diæt Restriktioner + Ingen Medicin Restriktioner + Nem at udføre + Lange screeningsintervaller	- Skal sendes til laboratorium - Svartid 2-3 uger - Dyr at udføre
Sigmoideoskopi	+ Høj specificitet + Høj sensitivitet + 80% af KRC opdages	- Invasiv - Blødning og perforering af tarm - Dyr at udføre - Kræver højtuddannet personale - Ubehageligt for patient
Koloskopi	+ Høj specificitet + Høj sensitivitet + >80% af KRC opdages + Lange screeningsintervaller	- Invasiv - Blødning og perforering af tarm - Dyr at udføre - Kræver højtuddannet personale - Ubehageligt for patient - Tyk og endetarmsrensning - Bedøvelse
Virtuel Koloskopi (CT)	+ Noninvasiv + Hurtig udførelse	- Meget dyr i anskaffelse - Manglende evidens - Tyk og endetarmsrensning - Dyrt udstyr
Klassifikator og datavarehus	+ Minimal invasiv + Hurtig og nem udførelse + Beslutningsstøtte til eg. FOBT + Høj sensitivitet + Genanvendelse af data	- Manglende evidens - Kræver kontakt - Kræver meget data - Højt antal falske positive

Tabel 7.1: *Opsummering på fordele og ulemper på de forskellige screeningsmetoder for KRC samt klassifikatoren til detektering af KRC med eksisterende data i datavarehuset.*

Flere andre datavarehusinitiativer har vist at datavarehuse kan bruges til flere forskellige kliniske formål [45, 41]. Som beskrevet tidligere har datavarehuset nogle begrænsninger ift. detektering af sygdomme, fordi der er begrænsede mængde data i datavarehuset før patienterne kommer i kontakt med sundhedsvæsnet. Men der er store mængder data til rådighed i datavarehuset omkring de enkelte kontakter, så der er potentiale for at bruge

datavarehuset til beslutningsstøtte, f.eks. ift. at analysere levetiden efter diagnosen er blevet stillet, eller hvilke procedurer, der giver den største chance for at patienten overlever KRC. I begge disse situationer vil det være interessant at implementere flere andre datakilder i datavarehuset, men specielt dødsårsagsregisteret er interessant.
Der er også potentiale for at anvende datamining ift. en anden type kræft eller vælge en helt tredje diagnose. Pointen er, at datavarehuset stiller store mængder data til rådighed, og at der er potentiale for at udvide datavarehuset med flere andre datakilder. Dermed bliver datavarehuset et stærk værktøj, som kan bruges til at lave initierende analyser ift. problemstillinger i sundhedssektoren. Datavarehuset kan blive et indledende skridt, der kan afgøre hvor ressourcerne til klinisk forskning bedste bruges.

Litteratur

[1] M. A. Boysen and J. B. Nielsen, "Dimensionel modellering af laboratoriedata og implementering af modellen i et datavarehus," projektrapport, Aalborg Universitet, Januar 2013.

[2] World Health Organization, "World health statistics 2012," 2012.

[3] F. Maffei, S. Angelini, G. C. Forti, and P. Hrelia, "Blood biomarkers linked to oxidative stress and chronic inflammation for risk assessment of colorectal neoplasia," Current Colorectal Cancer Reports, vol. 9, no. 1, pp. 85–94, 2013.

[4] T. V. Schroeder, S. Schulze, J. Hilsted, and L. Gøtzsche, Basisbog i Medicin og Kirurgi. Munksgaard Danmark, 4. udgave ed., 2010.

[5] Statens Serum Institut, "Sygehusbaseret kræftoverlevelse 1999-2010," 2013.

[6] Danmarks Statistik, "Tabel: Dod1 døde efter region, køn, alder og dødårsag." `http://www.statistikbanken.dk/statbank5a/SelectVarVal/Define.asp?MainTable=DOD1&PLanguage=0&PXSId=0&wsid=cftree`.

[7] Statens Serums Institut, "Nye kræfttilfælde fordelt på regioner." `http://www.ssi.dk/Sundhedsdataogit/Dataformidling/Sundhedsdata/Kraft/Nye%20krafttilfalde%20fordelt%20pa%20regioner.aspx`, 2013.

[8] B. Greenwald, "A comparison of three stool tests for colorectal cancer screening," MEDSURG Nursing, vol. 14, no. 5, pp. 292–300, 2005.

[9] M. Bretthauer, "Colorectal cancer screening," Journal of Internal Medicine, vol. 270, pp. 87–98, 2011.

[10] M. Schwab, ed., Encyclopedia of Cancer. Springer-Verlag Berlin Heidelberg, 3 ed., 2011.

[11] D. J. Kerr, A. M. Young, and F. D. R. Hobbs, ABC of Colorectal Cancer. BMJ Books, 2001.

[12] I. H. Clemmensen, K. H. Nedergaard, and H. H. Storm, Kræft I Danmark - En Opslagsbog. FADL's Forlag, September 2006.

[13] Sundhedsstyrelsen, "Kræftplan I - overblik." `http://www.sst.dk/Planlaegning%20og%20kvalitet/Kraeftbehandling/Nationale%20planer/Kraeftplan_1.aspx`, 2013.

[14] Sundhedsstyrelsen, "Kræftplan II - overblik." `http://www.sst.dk/Planlaegning%20og%20kvalitet/Kraeftbehandling/Nationale%20planer/Kraeftplan_II.aspx`, 2013.

[15] Sundhedsstyrelsen, "Kræftplan III - overblik." http://www.sst.dk/Planlaegning%20og%20kvalitet/Kraeftbehandling/Nationale%20planer/Kraeftplan_III.aspx, 2013.

[16] Sundhedsstyrelsen, "Aftale om kræftplan III," 2011.

[17] Sundhedsstyrelsen, "Anbefalinger vedrørende screening for tyk- og endetarmskræft." http://www.sst.dk/publ/Publ2012/09sep/ScreeningTarmkraeftAnbef2udg.pdf, 2012.

[18] Forskningscenter for Forebyggelse og Sundhed, Koncern Plan og Udvikling, and Region Hovedstaden, "Screening for tarmkræft i Vejle og Københavns Amter Tværgående evaluering af pilotprojekter." http://www.cancer.dk/NR/rdonlyres/39C14B4B-F99A-40AF-A14E-701A1C5F168E/0/Eksternevaluering.pdf, 2005.

[19] J. D. Hardcastle, J. O. Chamberlain, M. H. E. Robinson, S. M. Moss, S. S. Amar, T. W. Balfour, P. D. James, and C. M. Mangham, "Randomised controlled trial of faecal-occult-blood screening for colorectal cancer," The Lancet, vol. 348, no. November, pp. 1472–1477, 1996.

[20] I. Stegeman, T. R. de Wijkerslooth, E. M. Stoop, M. E. van Leerdam, E. Dekker, M. van Ballegooijen, E. J. Kuipers, P. Fockens, R. A. Kraaijenhagen, and P. M. Bossuyt, "Colorectal cancer risk factors in the detection of advanced adenoma and colorectal cancer," Cancer Epidemiology, vol. 537, 2013.

[21] S. Hundt, U. Haug, and H. Brenner, "Comparative evaluation of immunochemical fecal occult blood tests for colorectal adenoma detection," Annals of Internal Medicine, vol. 150, pp. 162–169, 2009.

[22] D. Karley, D. Gupta, and A. Tiwari, "Biomarkers: The future of medical science to detect cancer," J Mol Biomark Diagn, vol. 2, no. 5, pp. 1–7, 2011.

[23] W. Sujansky, "Heterogeneous database integration in biomedicine," Journal of Biomedical Informatics, vol. 34, pp. 285–298, 2001.

[24] M. Berg and E. Goorman, "The contextual nature of medical information," International Journal of Medical Informatics, vol. 56, pp. 51–60, 1999.

[25] Sundhedsstyrelsen, "Fælles udmøntningsplan for kræftplan III," 2011.

[26] Sundhedsstyrelsen, "Screening for tarmkræft - deltagelsesprocentens betydning – en medicinsk teknologivurdering." http://www.sst.dk/publ/Publ2008/MTV/screening_tarmkraeft/MTV_tarmkraeft_net_final_version2.pdf, 2008.

[27] S. Bulow, "Retningslinier for diagnostik og behandling af kolorektal cancer," 2009.

[28] R. A. Smith, V. Cokkinides, A. C. von Eschenbach, B. Levin, C. Cohen, C. D. Runowicz, S. Sener, D. Saslow, and H. J. Eyre, "American cancer society guidelines for the early detection of cancer," CA A Cancer Journal for Clinicians, vol. 52, pp. 8–22, 2002.

[29] B. Levin, D. A. Lieberman, B. McFarland, R. A. Smith, D. Brooks, K. S. Andrews, C. Dash, F. M. Giardiello, S. Glick, T. R. Levin, P. Pickhardt, D. K. Rex, A. Thorson, and S. J. Winawer, "Screening and surveillance for early detection of colorectal cancer and adenomatous polyps," CA A Cancer Journal for Clinicians, vol. 58, pp. 130–160, 2008.

[30] L. G. Van Rossum, A. F. Van Rijn, R. J. Laheij, M. G. Van Oijen, P. Fockens, H. H. Van Krieken, A. L. Verbeek, J. B. Jansen, and E. Dekker, "Random comparison of guaiac and immunochemical fecal occult blood tests for colorectal cancer in a screening population," Gastroenterology, vol. 135, pp. 82–90, 2008.

[31] A. G. Lalkhen and A. McCluskey, "Clinical tests: sensitivity and specifictiy," Continuing Education in Anaesthesia, Critical Care and Pain, vol. 8, no. 6, pp. 221–3, 2008.

[32] C. J. Kahi, J. C. Anderson, and D. K. Rex, "Screening and surveillance for colorectal cancer: state of the art," Gastrointestinal Endoscopy, vol. 77, no. 3, pp. 335–350, 2013.

[33] D. F. Ransohoff, "Colon cancer screening in 2005: Status and challenges," Gastroenterology, vol. 128, pp. 1685–1695, 2005.

[34] S. Hundt, U. Haug, and H. Brenner, "Blood markers for early detection of colorectal cancer: A systematic review," Cancer Epidemiol Biomarkers Prev, vol. 16, pp. 1935–1953, 2007.

[35] J. Y. Engwegen, H. H. Helgason, A. Cats, N. Harris, J. M. Bonfrer, J. H. Schellens, and J. H. Beijnen, "Identification of serum proteins discriminating colorectal cancer patients and healthy controls using surface-enhanced laser desorption ionisation-time of flight mass spectrometry," World J. Gastroenterol, vol. 12, pp. 1536–44, March 2006.

[36] A. Baerheim, "The diagnostic process in general practice: has it a two-phase structure?," Family Practice, vol. 18, pp. 243–245, 2001.

[37] T. P. Erlinger, E. A. Platz, N. Rifai, and K. J. Helzlsouer, "C-reactive protein and the risk of incident colorectal cancer," JAMA, vol. 291, no. 5, pp. 585–90, 2004.

[38] S. Sobrino-Cossio, E. Fenocchi, A. Hernandez-Guerrero, J. O. Alonso-Larraga, J. G. De la Mora-Levy, and I. Larracilla-Salazar, "Immunological fecal occult blood test vs. serum ferritin for detection of colorectal neoplasia in high risk asymptomatic population," Rev Gastroenterol Mex, vol. 76, no. 3, pp. 191–8, 2011.

[39] D. J. Berndt, J. W. Fisher, A. R. Hevner, and J. Studnicki, "Healthcare data warehousing and quality assurance," Computer, vol. 18, pp. 56–65, December 2001.

[40] H. U. Prokosch and T. Ganslandt, "Perspectives for medical informatics reusing the electronic medical record for clinical research," Methods Inf Med, vol. 48, pp. 38–44, 2009.

[41] R. S. Evans, J. F. Lloyd, and L. A. Pierce, "Clinical use of an enterprise data warehouse," AMIA Annu Symp Proc, pp. 198–198, 2012.

[42] A.-B. Wirehn, H. M. Karlsson, and J. M. Carstensen, "Estimating disease prevalence using a population-based administrative healthcare database," Scandinavian Journal of Public Health, vol. 35, pp. 424–431, 2007.

[43] M. F. Razavi, L. Falk, Å. Bjorn, and S. Wilhelmsson, "Experiences of the swedish healthcare system: An interview study with refugees in need of long-term health care," Scandinavian Journal of Public Health, vol. 39, pp. 319–325, March 2011.

[44] E. Heintz, A.-B. Wirehn, B. B. Peebo, L. Läns, U. Rosenqvist, and L. Levin, "Prevalence and healthcare costs of diabetic retinopathy: a population-based register study in sweden," Diabetologia, vol. 53, no. 10, pp. 2147–54, 2010.

[45] M. F. Wisniewski, P. Kieszkowski, B. M. Zagorski, W. E. Trick, M. Sommers, and R. A. Weinstein, "Development of a clinical data warehouse for hospital infection control," J Am Med Inform Assoc., vol. 10, pp. 454–462, 2003.

[46] R. Kimball and M. Ross, The Data Warehouse Toolkit. Wiley Publishing Inc, 2nd ed., 2002.

[47] M. Breslin, "Data warehousing battle of the giants: Comparing the basics of the kimball and inmon models," Business Intelligence Journal, pp. 6–20, Winter 2004.

[48] J. Ingernerf and W. Giere, "Concept-oriented standardization and statistics-oriented classification: Continuing the classification versus nomenclature controversy," Methods of Information in Medicine, vol. 37, pp. 527–39, 1998.

[49] E. Lynge, J. L. Sandegaard, and M. Rebolj, "The danish national patient register," Scandinavian Journal of Danish Registers, vol. 39, pp. 30–33, 2011.

[50] A. F. Grann, R. Erichsen, A. G. Nielsen, T. Frøslev, and R. W. Thomasen, "Existing data sources for clinical epidemiology: The clinical laboratory information system (LABKA) research database at aarhus university, denmark," Clinical Epidemiology, no. 3, pp. 133–138, 2011.

[51] R. A. Greenes, ed., Clinical Decision Support The Road Ahead. Academic Press, 2006.

[52] E. S. Berner, ed., Clinical Decision Support Systems. Springer New York, 2007.

[53] H. Chen, S. S. Fuller, C. Friedman, and W. Hersh, eds., Medical Informatics, vol. 8. Springer US, 2005.

[54] Institute for Work and Health, "What researchers mean by retrospective vs. prospective studies," At Work, vol. 59, p. 2, Winter 2010.

[55] N. J. Salkind, Encyclopedia of research design. SAGE Publications, inc., 2010.

[56] D. A. Ahlquist, "Occult blood screening - obstacles to effectiveness," Cancer Supplement, vol. 70, pp. 1259–1265, September 1992.

[57] D. Hand, "Statistical methods in diagnosis," Statistical Methods in Medical Research, vol. 1, pp. 49–67, 1992.

[58] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification. Wiley Publishing Inc, 2. ed ed., 2000.

[59] J. S. Mandel, J. H. Bond, T. R. Church, D. C. Snover, M. Bradley, L. M. Schuman, and F. Ederer, "Reducing mortality from colorectal cancer by screening for fecal occult blood," The New England Journal of Medicine, vol. 328, pp. 1365–1371, May 1993.

[60] A. K. Jain, R. P. Duin, and J. Mao, "Statistical pattern recognition: A review," IEEE Transactions On Pattern Analysis And Machine Intelligence, vol. 22, no. 1, pp. 4–37, 2000.

[61] M. H. Afif, A. Hedar, T. H. A. Hamid, and Y. B. Mahdy, "Ss-svm (3svm): A new classification method for hepatitis disease diagnosis," International Journal of Advanced Computer Science and Applications, vol. 4, no. 2, pp. 53–58, 2013.

[62] A. Indrayan, Basic Methods of Medical Research. A.I.T.B.S. Publishers, 2nd ed., 2008.

[63] C. J. Mann, "Observational research methods. research design II: cohort, cross sectional, and case-control studies," Emerg Med J, vol. 20, pp. 54–60, 2003.

[64] Sundhed.dk, "Lægehåndbogen - koagulationsfaktor 2,7,10 inr." https://www.sundhed.dk/sundhedsfaglig/laegehaandbogen/undersoegelser-og-proever/klinisk-biokemi/blodproever/inr-koagulation-vaevsfaktor-induceret-kf-ii-vii-x/, 2013.

[65] Sundhed.dk, "Lægehåndbogen - leukocytter." https://www.sundhed.dk/sundhedsfaglig/laegehaandbogen/undersoegelser-og-proever/klinisk-biokemi/blodproever/leukocytter-totale/, 2013.

[66] Sundhed.dk, "Lægehåndbogen - thrombocytter." https://www.sundhed.dk/sundhedsfaglig/laegehaandbogen/undersoegelser-og-proever/klinisk-biokemi/blodproever/trombocytter/, 2013.

[67] S. D. Zucker, P. S. Horn, and K. E. Sherman, "Serum bilirubin levels in the u.s. population: Effect and inverse correlation with colorectal gender cancer," Hepatology, vol. 40, no. 4, pp. 827–35, 2004.

[68] K. Diem and C. Lentner, Wissenschaftliche Tabellen: documenta Geigy. Georg Thieme, 1975.

[69] D. W. Kufe, R. E. Pollock, R. R. Weichselbaum, R. C. Bast, T. S. Gansler, J. F. Holland, and E. Frei, eds., Holland-Frei Cancer Medicine, vol. 2. BC Decker, 6th ed., 2003.

[70] Sundhedsstyrelsen, "Sks browseren - [dc15-dc26]." http://medinfo.dk/sks/brows.php?s_nod=6777, 2013.

[71] T. V. Schroeder and S. Schulze, Basisbog i Sygedomslære. Munksgaard Danmark, 2. udgave ed., 2011.

[72] C. C. Compton and F. L. Greene, "The staging of colorectal cancer: 2004 and beyond," CA: A Cancer Journal for Clinicians, vol. 54, pp. 295–308, 2004.

[73] C. C. Compton, D. R. Byrd, J. Garcia-Aguilar, S. H. Kurtzman, A. Olawaiye, and M. K. Washington, AJCC Cancer Staging Atlas 2012. Springer, 2. udgave ed., 2012.

Appendix

Baggrund A

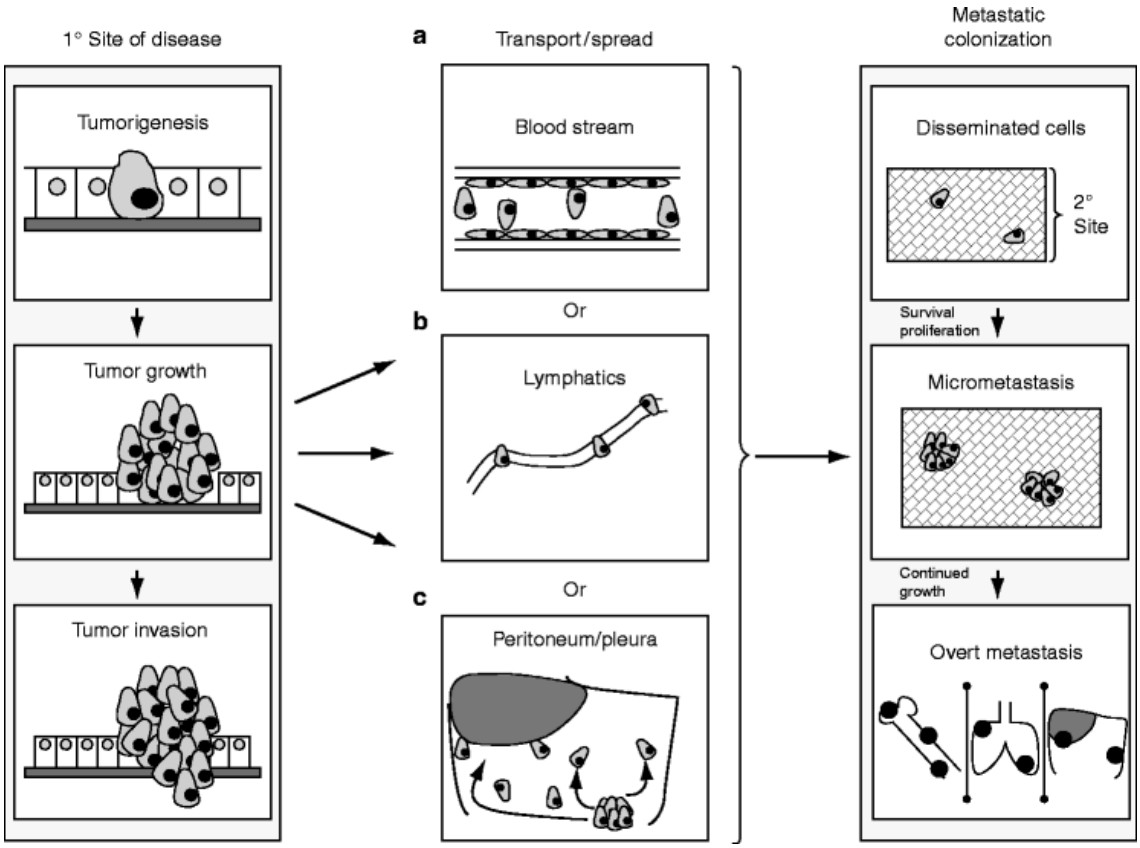
A.1 Onkologi

Kræft skyldes ondartet og anormal celledeling. Den anormale celledeling opstår ved en mutation i DNA'et, som er resultatet af skader på genet. Stråling, vira og kemiske stoffer er de bedst kendte faktorer, som resulterer i skader på genet. DNA'ets tendens til mutation kan også være arvelig[4, s. 737].

Anormal celledeling kaldes neoplasia, og det leder til dannelse af tumorer. Både godartede og ondartede tumorer er karakteriseret ved anormal vækst, hvor celledød sker langsommere end celledeling. En tumor er unik og heterogen, netop denne karakteristika er vigtig at huske, hvis man prøver at udvikle generaliserende modeller for kræft [4, s. 737]. Tumorerne kan være store eller små, og væksten kan være langsom eller hurtig. Fælles er dog at anormale celler akkumuleres [69, Kap 1]. Det indikerer altså, at der findes mange forskellige typer kræft. Kræftceller ligner de normale celler, hvor den første anormale celledeling er opstået. Kræft kan derfor klassificeres efter den type væv, som den første anormale celle kommer fra. Karcinoma er kræft opstået i epitelium, sarcoma er kræft opstået i bindevæv, lymphoma og leukemia stammer fra anormale hæmocytoblaster, som er de celler der udvikles til andre blodceller. Derudover findes kønscelletumorer, som opstår i de celler der producere kønsceller, samt blastoma, som er kræft, der er opstået fra præmature stamceller. I resten af dette afsnit vil fokus ligge på karcinoma, fordi de fleste tilfælde af kolorektal kræft (KRC) er af typen karcinoma.

En godartet tumor forbliver afgrænset af det epitelium, hvor den er opstået. En godartet tumor spreder sig ikke og vokser ikke ind i det omkringliggende væv. Den kan dog blive så stor, at den klemmer omkringliggende organer.
En ondartet tumor bliver også kaldt malign, hvilket betyder gradvis forværring med risko for død. En ondartet tumor forbliver ikke afgrænset af epitelium, den vokser ind i det omkringliggende væv og kan derved ødelægge væsentlige strukturer. Kræftcellerne som udgør tumoreren invaderer ikke kun omkringliggende strukturer, de angriber og dræber også normale celler. En ondartet tumor følger altså ikke reglerne for funktion og struktur, som andre celler gør[10, s. 605-607]. Der er oftest ingen symptomer eller indikationer på ondartede tumorer. Tidligt i forløbet bliver disse også afgrænset af epiteliet. Når symptomer opstår er disse knyttet til lokationen af tumoren. Ved. f.eks. kræft i hjernen kan symptomer som hovedpine og krampeanfald opleves. Disse symptomer kan ofte også kædes sammen med andre sygdomme. Kræft er derfor også blevet beskrevet som *den store efterligner* [10, s. 605-607]. Det kan være en bidragende faktor til at kræft ofte opdages sent i sygdomsforløbet.

En malign tumor er også karakteriseret ved evnen til at metastatisere. Dvs. at tumorceller river sig løs fra den primære tumor, og de spreder sig til andre lokationer i kroppen via blodbanen, lymfesystemet eller hulrum som bughulen [10, s. 2271-3]. På figur A.1 ses hvordan sekundære tumorer etableres væk fra den primære tumor. Metastasis gør, at kræft bliver en kompleks sygdom, som involverer hele den menneskelige organisme[10, s. 2260-1]. Før metastisering kan ske, skal tumoren udvise invasiv karakter. Først da kan tumorcellerne nå lymfeknuderne eller en af de andre transporteringsmuligheder, som er illustreret ved a,b,c på figur A.1.



Figur A.1: *Primær tumoren udvikler sig og invaderer det omkringliggende væv. Dernæst spreder den sig via blodbanen, lymfesystemet eller hulrum. Dette leder til metastase, altså spredning af kræftceller til andre dele af kroppen. Her prøver kræftcellerne at overleve og dele sig. Denne proces kan være langvarig, og ofte er metastaserne inaktive i længere perioder, før metastaserne bliver klinisk åbenlyse [10, s. 2260-1].*

Udvikling af kræft er en darvinistisk proces, hvor almindelig celler muterer til kræftceller ved anormal celledeling. Disse kræftceller overholder ikke de normale regler for celledeling, og de vil derfor fortrænge og bekæmpe de normale celler[10, s. 2260-1].
Genetiske analyser har identificeret en række gener, som sammen får kræft til at opstå. Disse gener er samlet i seks overordnede kategorier. Nogle kræfttyper udviser kun mutationer i nogle af kategorierne, mens andre typer udviser mutationer i næsten alle kategorierne. Kategorierne er:

1. Aktivering af et onkogen (gen der har potentiale til udvikle kræft), hvilket får cellen til øge replikationshastigheden.
2. Inaktivering af tumorsuppressorgen. Tumorsuppressorgen stopper anormal celledeling ved at styre cellereplikationshastighed og celledød.
3. Aktivering af gener, der øger cellens levetid, altså skaber en ubalance mellem celledeling og celledød.
4. Inaktivering af DNA-repareringsgener, hvilket skaber ustabilitet i genomet.
5. Inaktivering af gener, der sætter en grænse for antallet af celledelinger, hvilket vil sige at en celle kan dele sig et ubestemt antal gange, det siges også at celler uden disse gener er udødelige.
6. Gener der har indflydelse på metastisering og invasiv vækst.

Generelt finder 1, 2 og 3 sted for at en normal celle udvikler sig til at kræftcelle[10, s 2405-8]. Mutationer i disse kategorier kan opstå på forskellige tidspunkter og i forskellig rækkefølge. Det kan forklare, hvorfor nogle kræfttyper kan være mere end 20 år om at udvikle metastaser, som er den sidste og ofte fatale fase af kræft.

A.2 Kolorektal Kræft

Kolorektal kræft (KRC) omfatter kræft i tyktarmen, blindtarmen, og endetarmen. Dette er i SKS beskrevet ved komplekserne [DC18] og [DC20], som indeholder de underliggende SKS koder for hver af de forskellige lokationer af kræften: [DC181-9], [DC209-X] [70].

KRC menes at opstå ved, at en celle i epitelet skaber forandringer, som over tid spredes til slimhinden, og senere en dybere spredning til det omkringliggende væv. Bortset fra viden om, at det er lokale celleforandringer, der udløser kræften, er KRC's ætiologi delvis ukendt. Undersøgelser og forskning viser, at både arvelige årsager såvel som miljømæssige faktorer kan være udløsende faktorer for de celleforandringer, der finder sted i epiteliet. Der er også fundet en sammenhæng mellem KRC og livsstil, hvor overvægt, manglende motion, kost med høj koncentration af animalsk fedt og lavt fiberindhold kan være medvirkende til KRC. Desuden kan inflammatoriske tarmsygdomme så som Morbus Chrons også øge risikoen for KRC [4, s. 143][11, s.2-3][3].

A.2.1 Symptomer, Undersøgelser og Fund

Oftest er symptomerne på KRC blod i afføring, vægttab og smerter. Men disse symptomer manifesterer sig ikke i alle tilfælde, hvorfor undersøgelse for og diagnosticering af KRC sker gennem rektaleksploration. Af andre symptomer er der desuden ændrede afføringsvaner - skiftevis obstipation og diare, træthed og madlede [71, s. 54]. KRC kan også lede til anæmi især ved højresidige tumorer.

Undersøgelser for KRC foregår som nævnt gennem rektaleksploration i form af enten sigmoideoskopi eller koloskopi (endoskopi). Endoskopi udføres hovedsageligt på alle patienter over 40 år, hvor der er mistanke om KRC. Denne aldersgrænse skyldes, at op imod 99% af alle patienter, der diagnosticeres med KRC, er over 40 år [6, 7][4, s. 143]. I forbindelse med at der gennemføres endoskopi, tages ligeledes biopsi af mistænkte knuder og adenomer. I tilfælde hvor patienten ikke ønsker at få foretaget endoskopi, eller det ikke er muligt, tages

røntgenbilleder af abdomen og tarme. Der findes endnu ikke nogen påvist sammenhæng mellem blodprøver og KRC, som med sikkerhed kan bruges til at diagnosticere tilstedeværelsen af KRC[4].
Ved undersøgelse for KRC kan 2/3 af tilfældene palperes ved rektaleksploration. Ved større tumorer kan disse palperes abdominalt, dette er hyppigst ved de højresidige tumorer, da disse oftest opdages sent. Desuden kan der i nogle tilfælde ses forstørret lever på grund af metasaser. KRC metastaserer via lymfekar, blodkar og peritoneum til andre organer, hyppigst til leveren (op mod 50 %) og dernæst lunger[4, s. 145-146].
Under diagnosticeringen, hvor kræften skal klassificeres og stadieinddeles, anvendes MR eller ultralydsscanning. Dette bruges også i forhold til præoperativ strålebehandling. Som en del af udredningen foretages desuden ultralydsscanning af leveren for at konstatere eventuel metastasering [4, s. 145-146].

A.2.2 Klassifikation vha. TNM systemet

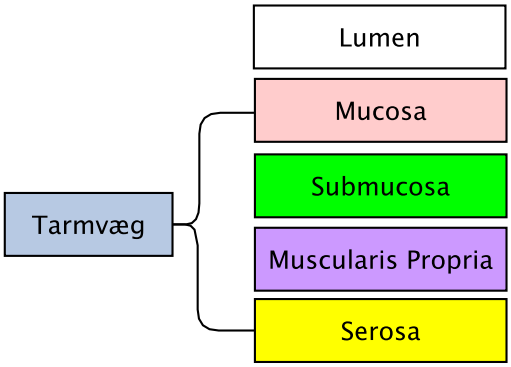
KRC inddeles i flere forskellige stadier alt efter, hvor meget kræften har spredt sig. I dag anvendes TNM systemet til stadieinddeling og klassifikation af KRC[11, s.8].

TNM systemet er et internationalt anvendt klassifikationssystem for kræft, og indeholder både kliniske og patologiske stadieinddelingsmetoder. TNM systemet udvikles og tilpasses løbende i takt med, at der kommer ny viden omkring de forskellige kræftformer, sådan at de prognoser, der stilles for hvert af stadierne, passer bedst muligt [72].

I hvert tilfælde vurderes tre områder, tumor (T), lymfeknuder (N), og metastaser (M).
T beskriver omfanget af den ubehandlede primære tumor. N beskriver spredningen af kræft til de regionale lymfeknuder og M beskriver fjernmetastaser på sygdomstidspunktet. T,N og M værdierne angives med enten et p for patologisk eller c for klinisk når kræften vurderes. Vurderes kræften inden, der er indledt nogen form for behandling, tilføjes c foran T,N og M, da der på dette tidspunkt sjældent er patologiske data til rådighed. Efter iværksat behandling bruges p oftest som præfiks. Et eksempel på en vurdering kan være pT1, pN0, cM0. Det er kun en samlet vurdering af både T, N og M som kan give et stadie for kræften. T kan vurderes til T*is*: Carcinoma in situ. Det er et forstadie til en tumor, hvor cellerne ligner kræftceller, men de er ikke begyndt at penetrere vævet. T1: Tumor invaderer submucosa. T2: Tumor invaderer Muscularis propria. T3: Tumor invaderer Muscularis propria og ind i serosa. T4a: Tumoren penetrerer serosa og vokser ud af tarmen. T4b: Tumoren penetrerer omkringliggende væv efter at have penetreret serosa. Figur A.2 viser organisering af de vævslag, som blev nævnt i klassificeringen af primære tumor(T). N kan vurderes til N0: Ingen metastaser i regionale lymfeknuder. N1: Metastase i 1-3 lymfeknuder. N1c: Ingen metastase i regionale lymfeknuder, men der findes tumorceller, som er adskilt fra, men lokaliseret tæt på primær tumoren. N2: Metastase i 4 eller flere lymfeknuder (N2a:4-6, N2b:7 og derover). M kan vurderes til M0: Ingen fjern metastaser. M1: Fjern metastaser (M1a: Et organ, M1b flere organer/bughulen) [72][73, s. 191-192].

På baggrund af den samlede TNM vurdering gives et stadie for alvorligheden af KRC. Stadiernes relation til TNM systemet er beskrevet i tabel A.1.

Fra tabel A.1 ses det, at der grundlæggende er fire forskellige stadier, da 0 indikerer, at der ikke er nogen kræftsvulster. I forstadierne 2-4 er der tillige en underinddeling af stadierne. Men den



Figur A.2: *Organisering af vævslag i tarmen. Lumen er hulrummet i tarmen, og Serosa er ydersiden*

Stadie	TNM Kombination		
0	Tis	N0	M0
1	T1-T2	N0	M0
2A	T3	N0	M0
2B	T4a	N0	M0
2C	T4b	N0	M0
3A	T1-T2	N1/N1c	M0
	T1	N2a	M0
3B	T3-T4a	N1/N1c	M0
	T2-T3	N2a	M0
	T1-T2	N2b	M0
3C	T4a	N2a	M0
	T3-T4a	N2b	M0
	T4b	N1-N2	M0
4a	Tis-T4b	N1-N2b	M1a
4b	Ethvert T	Ethver N	M1b

Tabel A.1: *Tabel over de forskellige stadier i KRC med tilhørende TNM klassifikation [72][73, s. 185-192].*

overordnede beskrivelse for stadierne 1-4 er som beskrevet herunder [11, s. 8][73, s. 185-192]:

- Stadie 1 - Kræften er ikke vokset igennem tarmvæggen, og der er ingen spredning til lymfeknuder eller andre metastaser.
- Stadie 2 - Kræften er vokset igennem tarmvæggen, og evt. ind i andre organer. Der er ingen spredning til lymfeknuder eller andre metastaser.
- Stadie 3 - Der er spredning til de regionale lymfeknuder.
- Stadie 4 - Der er fjernmetastaser.

A.2.3 Behandling og Prognose

Den normale behandling for KRC er præoperativ strålebehandling efterfulgt af kirurgisk fjernelse af kræftvæv (resektion). Det gælder tillige for metasaser, hvis det er muligt. Oftest iværksættes en onkologisk behandling af kræften. 75% af patienterne kan behandles radikalt ved fjernelse af synligt tumorvæv. Overlevelsesprognosen er tæt relateret til stadieinddelingen. Den gennemsnitlige 5 årige overlevelse er på ca. 50%, men mange dør af sekundære årsager [11, s.8]. Den gennemsnitlige 5 årige overlevelse kan dog være så høj som 90%, hvis kræften behandles tidligt i forløbet [10, s. 938]. Studier har desuden vist, at dødeligheden for KRC kan mindskes ved at fjerne adenomatøse polypper[3]. De forskellige 5 årige relative overlevelseschancer for hvert af de fire stadier er [11, s. 8]:

- Stadie 1 - 90-100% overlevelse.
- Stadie 2 - 75-85% overlevelse.
- Stadie 3 - 30-40% overlevelse.
- Stadie 4 - < 5% overlevelse.

I forbindelse med behandlingen optræder, der i 1/4 af tilfældene komplikationer i form af perforering af tarmvæggen. 20% af patienterne har risiko for utæthed i sammensyningen af tarmen i forbindelse med rectumresektion – dødeligheden for patienter, der har utæthed i sammensyningen af tarmen, er på 30%. I forbindelse med kolonresektion er der kun en risiko på 2-4% for utæthed i sammensyningen[71, s. 53-55].

FaktaCenter - et dimensionelt datavarehus

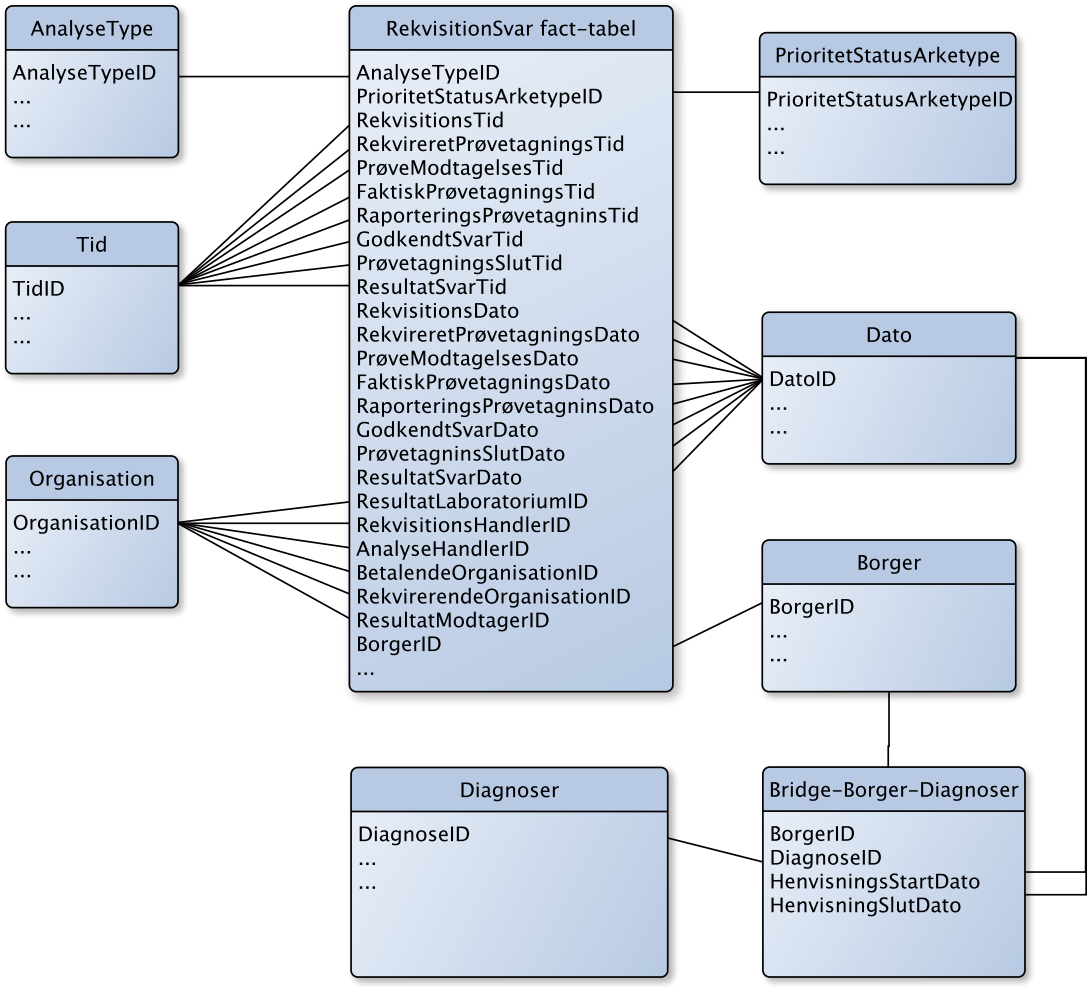
B

FaktaCenterets underliggende datamodel følger det paradigme, der hedder dimensionel modellering. Dvs. at FaktaCenteret ikke indeholder en komplet model af alle data i Region Nordjylland. Det indeholder en model, som langsomt kan udvides med flere og flere data.
Et dimensionelt datavarehus er i den simpleste forstand en database, som kan bruges til at integrere data fra flere forskellige datakilder og dermed klargøre data til sekundær brug. Dimensionelle datevarehuse har en række vigtige egenskaber[46]:

- Det skaber nem tilgang til data fra mange datakilder. Nem tilgang betyder ikke kun hurtig fysisk adgang til data, men også at data er forståelige.
- Datavarehuset skal også kunne præsentere konsistente data, altså kun en version skal være tilgængelig. Hvis ikke parametre, som måler det samme også hedder det samme, så bliver dataudtrækkene inkonsistente.
- Den dimensionelle model kan udvides. Det gøres med én arbejdsproces ad gangen. F.eks. hovedarbejdsprocessen på et laboratorium, som starter med en rekvirering af prøven og slutter med levering af resultatet.
- Integrering af data vha. konforme dimensioner.

Et dimensionelt datavarehus består af dimensioner og fact-tabeller.
En dimension er en samling af beskrivende attributter. Disse attributter er med til at gøre fact-tabellen forståelig. Attributter bruges altså som etiketter i rapporter og til at filtrere forespørgsler. Fact-tabellen er en tabel, hvis attributter er aggregerbare, og en række repræsenterer en eller anden form for transaktion. F.eks. oprettes en række i fact-tabellen, hver gang en blodprøve analyseres. Oftest er disse attributter additive, fordi man næsten altid aggregerer flere rækker. Attributter til fact-tabellen bruges ofte til beregninger, og de ændres ofte hurtigt, det betyder at antallet af rækker ofte er meget stort. Dimensioner er ofte denormaliserede, og det betyder at dimensioner bliver meget brede men relativt korte.
Den vigtigste styrke i en dimensionel model er de konforme dimensioner. En konform dimension er en dimension, der bruges af flere fact-tabeller. F.eks. bruges borgerdimensionen både af fact-tabeller tilknyttet PAS og fact-tabeller tilknyttet LABKAII. Det er ved brugen af konforme dimensioner, at data kan integreres i forespørgsler. Dermed sikres en fælles og konsistente datagrundlag.

I fact-tabellen findes også surrogatnøgler til de dimensioner, som fact-tabellen er tilknyttet. Figur B.1 viser et stjernediagram, som illustrerer hvordan fact-tabellen indeholder surrogatnøgler, der skaber forbindelsen til dimensionernes forretningsnøgle.



Figur B.1: *Stjernediagrammet illustrer den dimensionelle model af LABKAII, som er implementeret i FaktaCenteret. De omkringliggende dimensioner er tilknytter fact-tabellen, der ligger i centrum af figuren. Nederst i figuren er bridgetabel tilføjet for at sikre integrering mellem LABKAII-modellen og diagnosekoderne i PAS-modellen [1]*

B.1 Integreringen mellem laboratoriedata og patientadministrative data

Dette projekt inkluderer en integreringsopgave i og med målet er at undersøge laboratorieprøver fra patienter med kolorektal kræft (KRC). Der er behov for at tilknytte en diagnose til de patienter, der har fået tage en blodprøve.
For at lave integreringen mellem laboratoriedata og diagnoser har vi oprettet en bridge-tabel. En bridge-tabel er det man kalder en *factless* fact-tabel, fordi der ingen parametre er i den.

Den indeholder kun ID'er. Bridge-tabellen er dannet ved at finde start og slut på en kontakt samt patienter og alle diagnoserne tilknyttet den specifikke kontakt. En kontakt kan have mange diagnoser tilknyttet. F.eks. vil en kontakt med to diagnoser tilknyttet resultere i 2 rækker i bridge-tabellen. Fact-tabellen med laboratoriedata indeholder et patient id samt datoer på bestillingen af analyserne. For at koble diagnosekoder på de enkelte analyser, skal man blot flette de to tabeller med patientid'et og datoerne (se figur B.1). Selvom bridge-tabellen indeholder 22 millioner rækker, så effektiviserer den forespørgslen, hvor en diagnosekode skal kobles på en patient i LABKAII, fordi man kun skal henover denne ene bridgetabel i stedet for at skulle henover den store kontakt fact-tabel i PAS-modellen, som er væsentlig bredere.

Klassifikatorer, ROC Kurver og Beslutningsplot

C

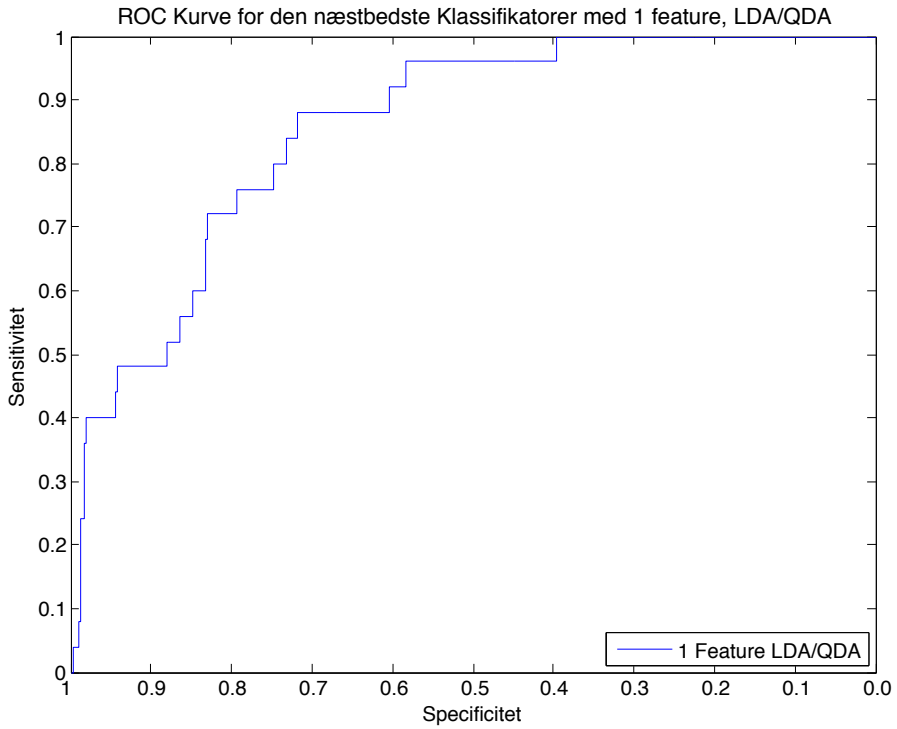
C.1 Klassifikatorer baseret på én feature

2. Klassifikator	**LDA**	**QDA**
Feature kombination	L.Reg. P-Natriumion	L.Reg. P-Natriumion
Positiv Prædiktivitet [%]	6,2	6,2
Negativ Prædiktivitet [%]	99,7	99,7
Sensitivitet [%]	88,0	88,0
Specificitet [%]	71,8	71,8
Fejlrate [%]	27,9	27,9

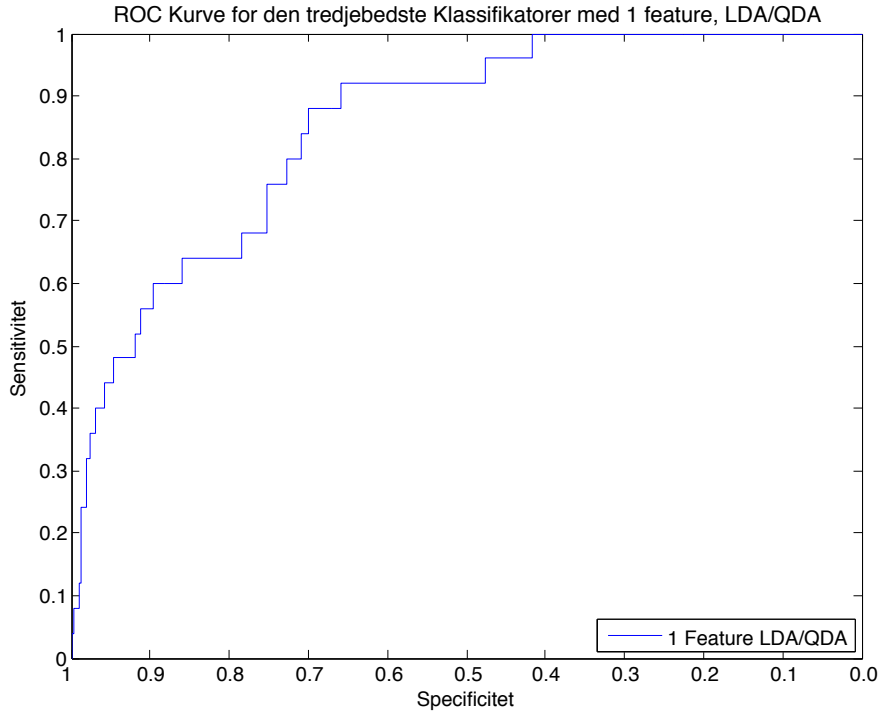
3. Klassifikator	**LDA**	**QDA**
Feature kombination	L.Reg. B-Hæmoglobin	L.Reg. B-Hæmoglobin
Positiv Prædiktivitet [%]	5,8	5,8
Negativ Prædiktivitet [%]	99,6	99,6
Sensitivitet [%]	88,0	88,0
Specificitet [%]	69,9	69,9
Fejlrate [%]	29,7	29,7

Valideringssamples	
Antal KRC	25
Antal Kontrol	1186

Tabel C.1: *De næst- og tredjebedste kassifikatorer med højest sensitivitet ift. specificitet for henholdvis LDA og QDA baseret på én feature*



Figur C.1: *Figuren viser ROC kurven for de klassifikatorer med næst højeste sensitivitet og specificitet baseret på én feature. Resultatet er vist i tabel C.1 for 2. Klassifikator*



Figur C.2: *Figuren viser ROC kurven for de klassifikatorer med tredje højeste sensitivitet og specificitet baseret på én feature. Resultatet er vist i tabel C.1 for 3. Klassifikator*

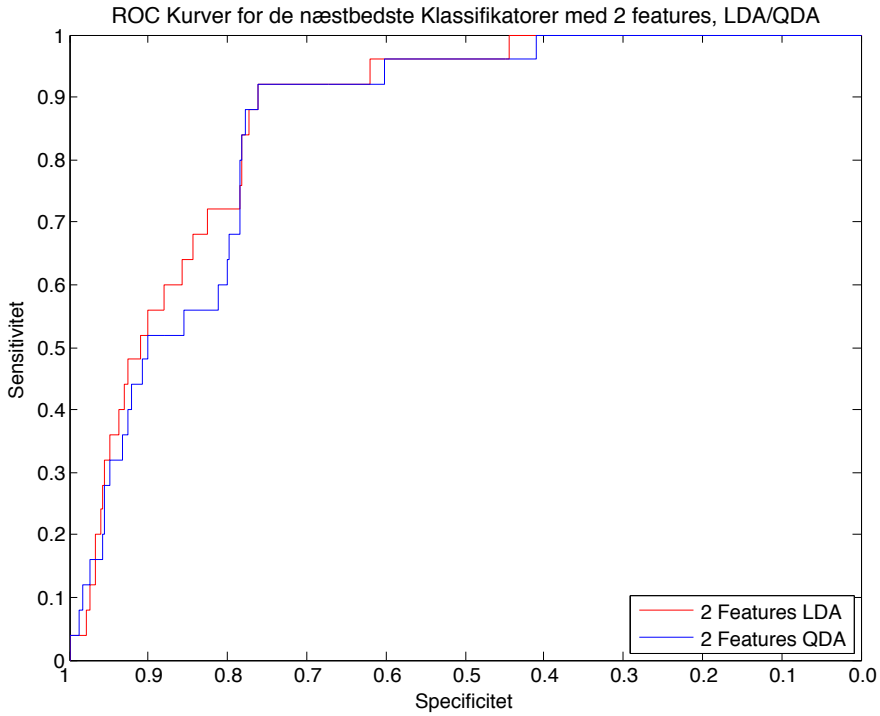
C.2 Klassifikatorer baseret på to features

2. Klassifikator	**LDA**	**QDA**
Feature kombination	L.Reg. B-Leukocytter L.Reg. P-Albumin MCV	Gns. B-Erythrocytter MCV L.Reg. P-Albumin
Positiv Prædiktivitet [%]	7,5	7,5
Negativ Prædiktivitet [%]	99,8	99,8
Sensitivitet [%]	92,0	92,0
Specificitet [%]	76,2	76,1
Fejlrate [%]	23,5	23,5

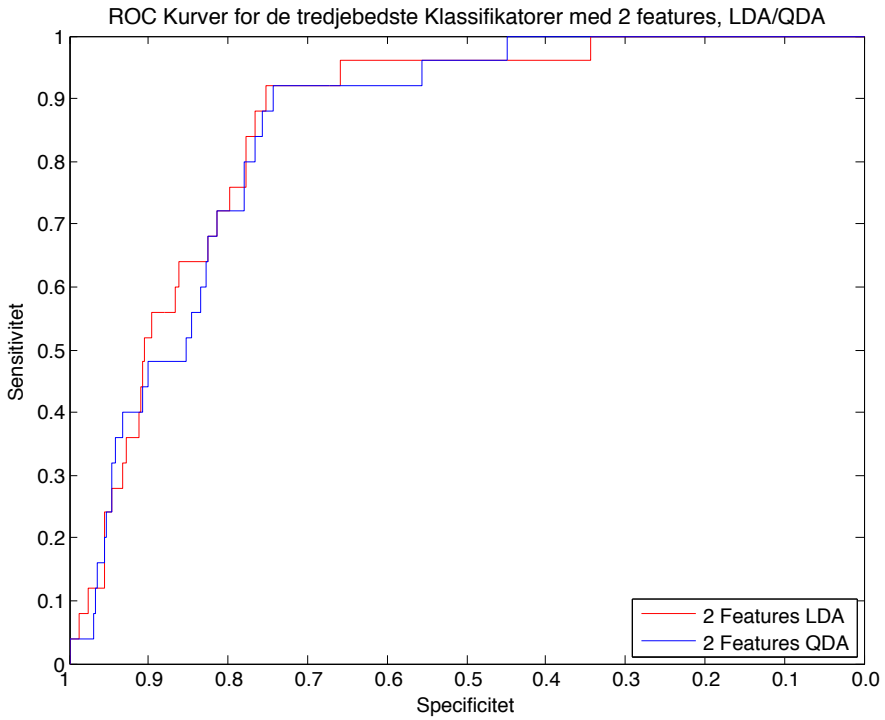
3. Klassifikator	**LDA**	**QDA**
Feature kombination	Std. P-Koagulationsfaktorer L.Reg. P-Albumin	Std. B-Thrombocytter L.Reg. P-Albumin
Positiv Prædiktivitet [%]	7,3	7,0
Negativ Prædiktivitet [%]	99,8	99,8
Sensitivitet [%]	92,0	92,0
Specificitet [%]	75,3	74,4
Fejlrate [%]	24,4	25,3

Valideringssamples	
Antal KRC	25
Antal Kontrol	1186

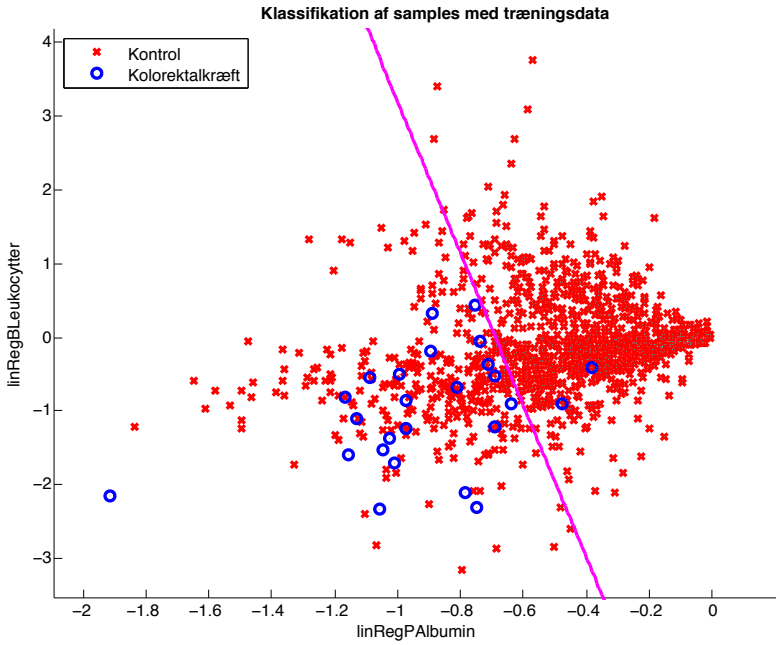
Tabel C.2: *De næst- og tredjebedste klassifikatorer med højest sensitivitet ift. specificitet for henholdvis LDA og QDA baseret på to features*



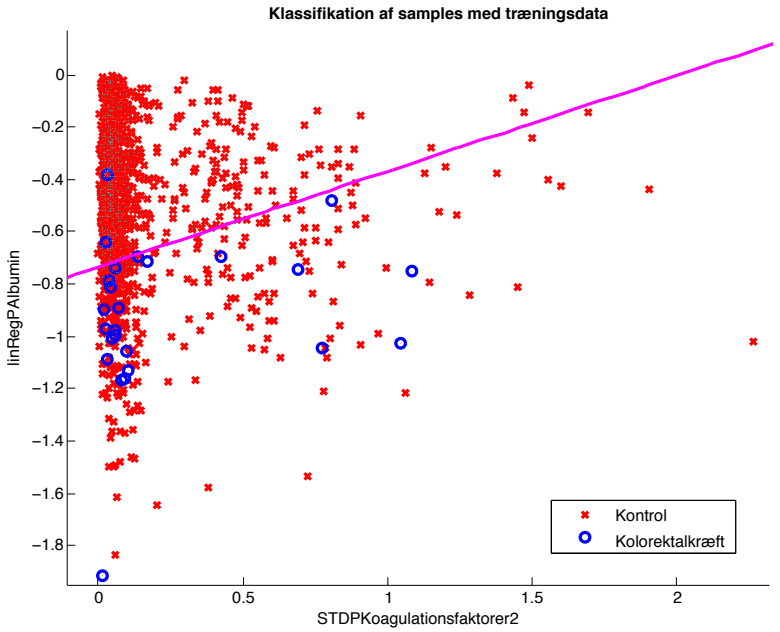
Figur C.3: *Figuren viser ROC kurven for de klassifikatorer med næsthøjest sensitivitet og specificitet baseret på to features for hhv. LDA og QDA. Resultatet er vist i tabel C.2 for 2. Klassifikator*



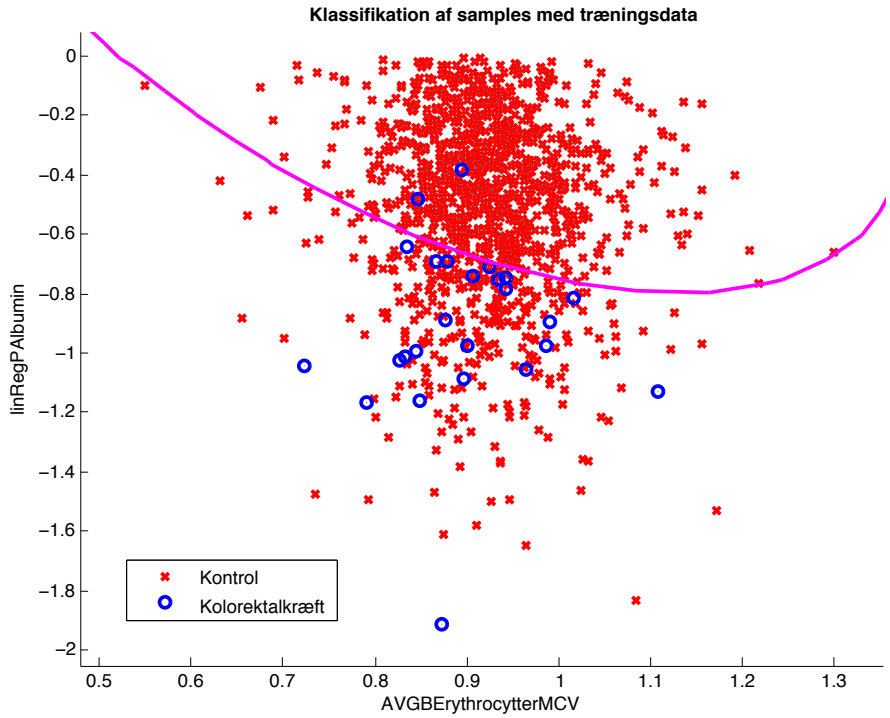
Figur C.4: *Figuren viser ROC kurven for de klassifikatorer med tredjehøjest sensitivitet og specificitet baseret på to features for hhv. LDA og QDA. Resultatet er vist i tabel C.2 for 3. Klassifikator*



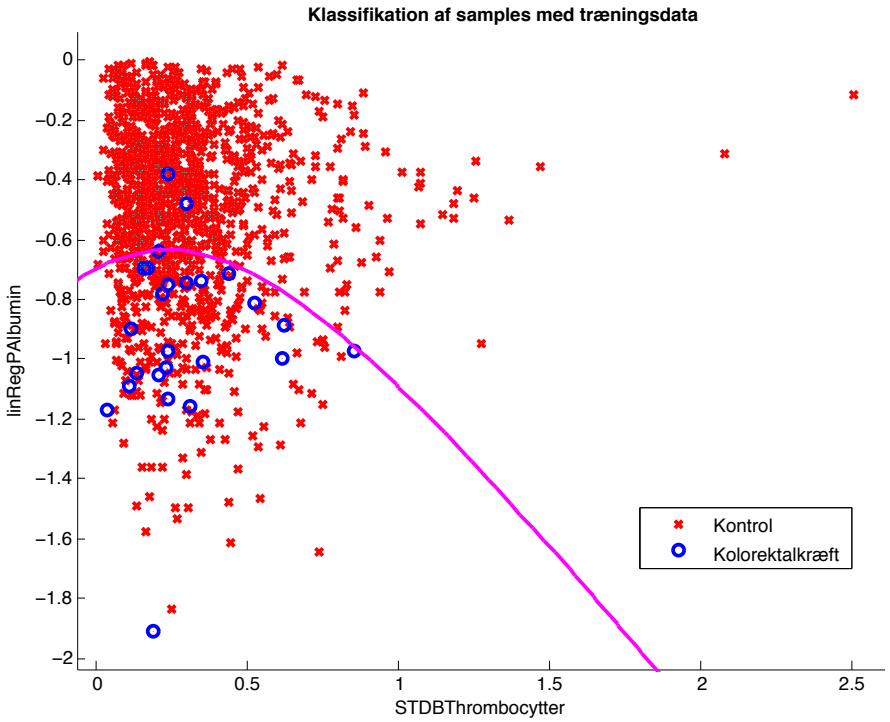
Figur C.5: *Figuren viser beslutningslinjen for den lineære klassifikator med næsthøjest sensitivitet og specificitet baseret på to features. Jvf. tabel C.2 for 2. Klassifikator LDA.*



Figur C.6: *Figuren viser beslutningslinjen for den lineære klassifikator med tredjehøjeste sensitivitet og specificitet baseret på to features. Jvf. tabel C.2 for 3. Klassifikator LDA.*



Figur C.7: *Figuren viser beslutningslinjen for den kvadratiske klassifikator med næsthøjeste sensitivitet og specificitet, baseret på to features. Jvf. tabel C.2 for 2. Klassifikator QDA.*



Figur C.8: *Figuren viser beslutningslinjen for den kvadratiske klassifikator med tredjehøjeste sensitivitet og specificitet, baseret på to features. Jvf. tabel C.2 for 3. Klassifikator QDA.*

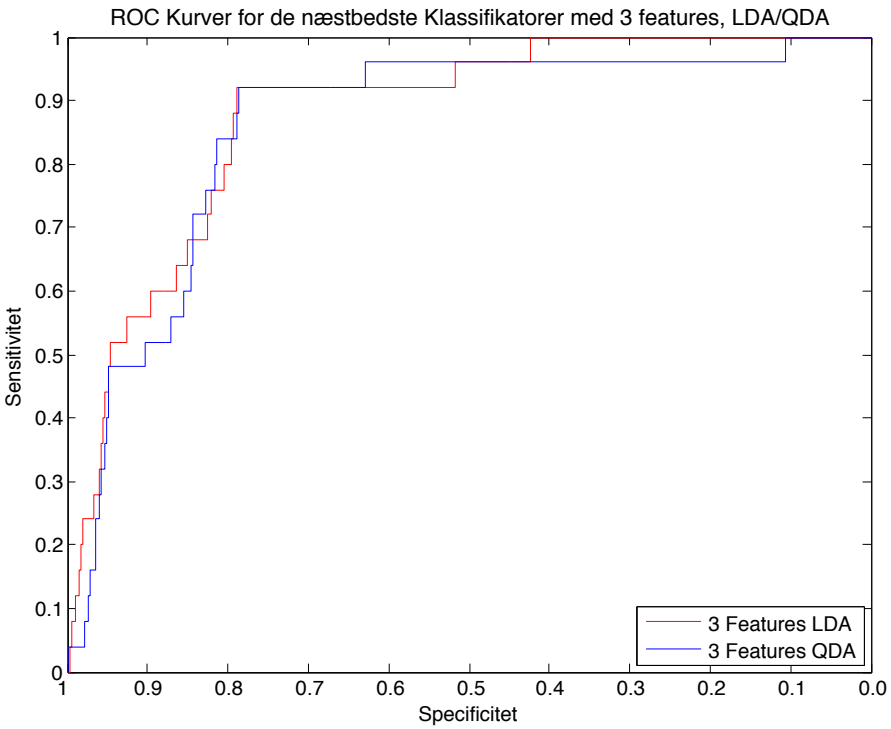
C.3 Klassifikatorer baseret på tre features

2. Klassifikator	**LDA**	**QDA**
Feature kombination	Gns. P-Koagulationsfaktorer L.Reg. P-Natriumion L.Reg. B-Leukocytter	Gns. P-Calcium total Gns. B-Thrombocytter L.Reg. P-Albumin
Positiv Prædiktivitet [%]	8,4	8,3
Negativ Prædiktivitet [%]	99,8	99,8
Sensitivitet [%]	92,0	92,0
Specificitet [%]	78,9	78,7
Fejlrate [%]	20,8	21,1

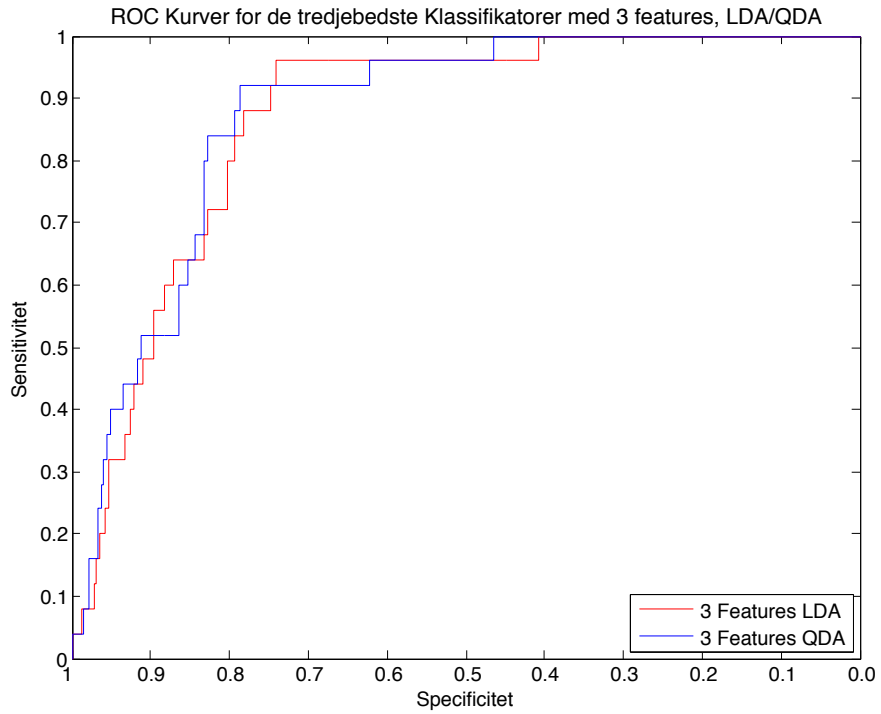
3. Klassifikator	**LDA**	**QDA**
Feature kombination	Gns. P-Koagulationsfaktorer L.Reg. P-Albumin L.Reg. B-Leukocytter	Std. Creaktivt Protein CRP Gns. B-Thrombocytter L.Reg. P-Albumin
Positiv Prædiktivitet [%]	7,3	8,3
Negativ Prædiktivitet [%]	99,9	99,8
Sensitivitet [%]	96,0	92,0
Specificitet [%]	74,2	78,6
Fejlrate [%]	25,4	21,1

Valideringssamples	
Antal KRC	25
Antal Kontrol	1186

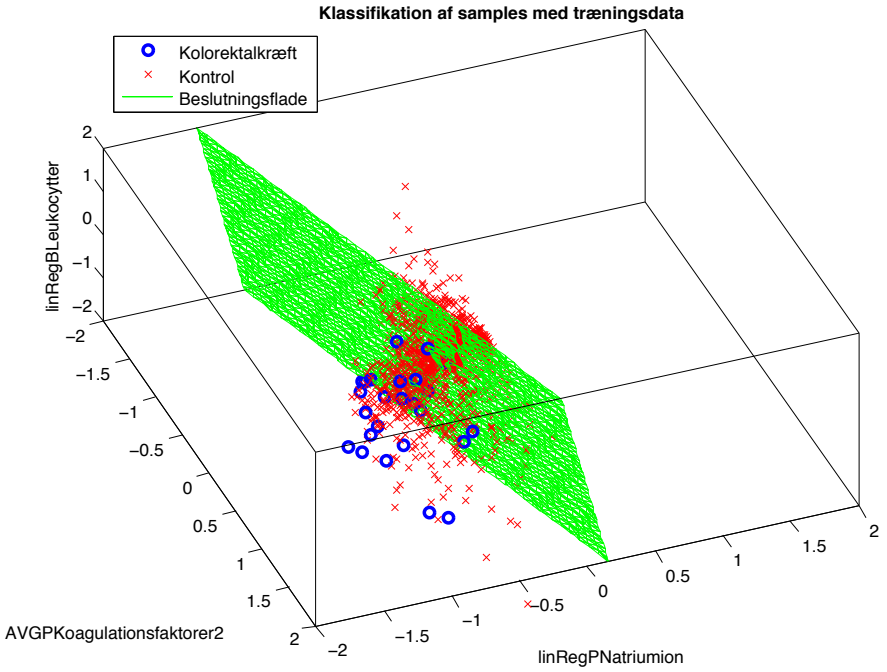
Tabel C.3: *De næst- og tredjebedste klassifikatorer med højest sensitivitet ift. specificitet for henholdvis LDA og QDA baseret på tre features*



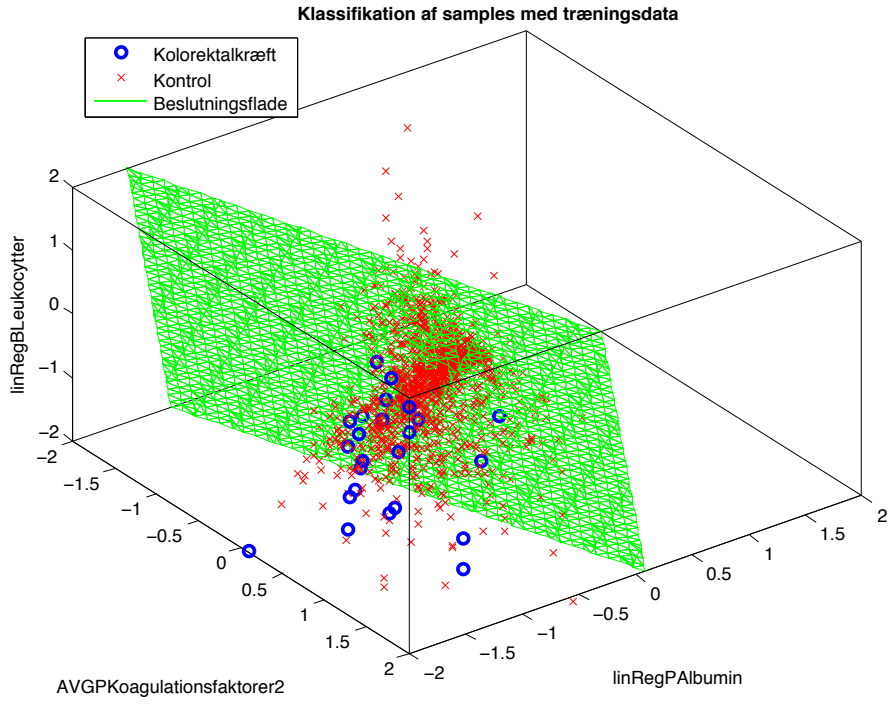
Figur C.9: *Figuren viser ROC kurven for de klassifikatorer med næsthøjeste sensitivitet og specificitet baseret på tre features, for hhv. LDA og QDA. Resultatet er vist i tabel C.3 for 2. Klassifikator*



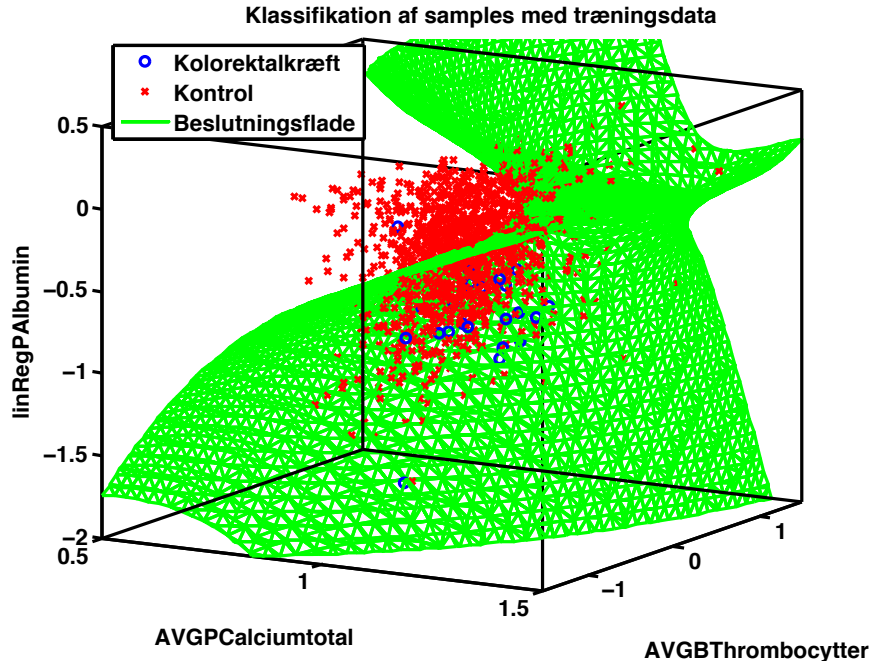
Figur C.10: *Figuren viser ROC kurven for de klassifikatorer med tredjehøjeste sensitivitet og specificitet baseret på tre features, for hhv. LDA og QDA. Resultatet er vist i tabel C.3 3. Klassifikator*



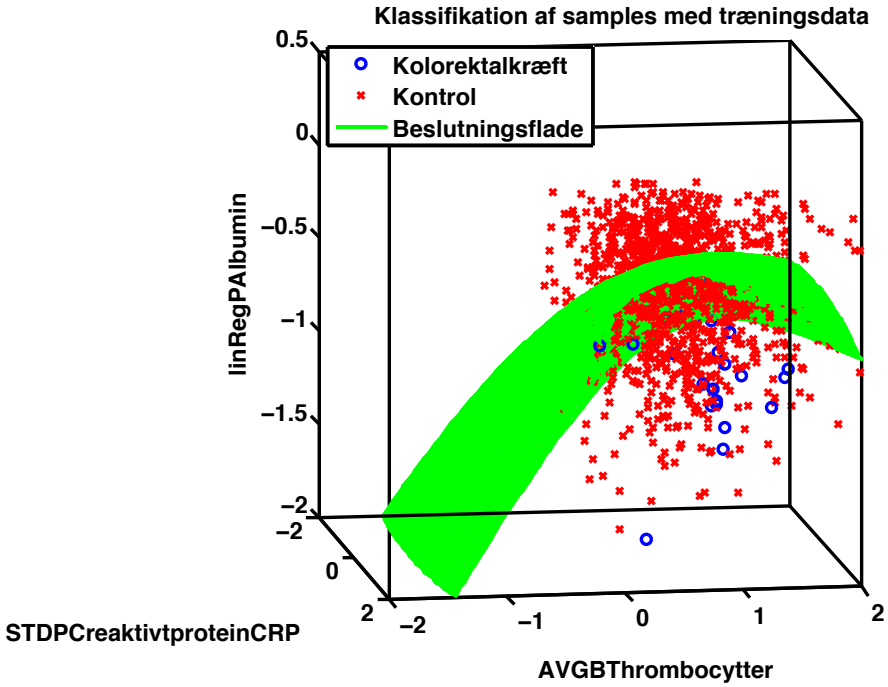
Figur C.11: *Figuren viser beslutningslinjen for den lineære klassifikator med næsthøjeste sensitivitet og specificitet, baseret på tre features. Jvf. tabel C.3 for 2. Klassifikator LDA.*



Figur C.12: *Figuren viser beslutningslinjen for den lineære klassifikator med tredjehøjeste sensitivitet og specificitet baseret på tre features. Jvf. tabel C.3 for 3. Klassifikator LDA.*



Figur C.13: *Figuren viser beslutningslinjen for den kvadratiske klassifikator med næsthøjeste sensitivitet og specificitet baseret på tre features. Jvf. tabel C.3 for 2. Klassifikator QDA.*



Figur C.14: *Figuren viser beslutningslinjen for den kvadratiske klassifikator med tredjehøjeste sensitivitet og specificitet baseret på tre features. Jvf. tabel C.3 for 2. Klassifikator QDA.*