Identifikation af markører for vigtige behandlingsudfald i elektroniske patientjournaler for patienter på Psykiatrisk Center Sct. Hans med psykiatriske diagnoser hvor behandling med antipsykotika er indiceret

Kombinationsspeciale i datalogi og medicinalbiologi

Af Kasper Klitgaard

Roskilde Universitet

Forskningsinstituttet, Psykiatrisk Center Sct. Hans

Vejledere:
Henrik Bulskov (Datalogi, RUC)
Ole Vang (Medicinalbiologi, RUC)
Thomas Werge (Forskningsinstituttet, Psykiatrisk Center Sct. Hans)

Resumé

I løbet af indlæggelsen på Psykiatrisk Center Sct. Hans indsamles informationer fra behandlingsforløbet i den elektroniske patientjournal (EPJ). Ud fra en hypotese om, at udvalgte parametre blandt de journalførte informationer udgør markører for det pågældende behandlingsudfald, forsøges disse markører fastlagt i et retrospektivt kohorte-studie af 354 psykiatriske patienter med behandlingsudfaldene udskrivning ($n = 224$) og behandlingsskifte ($n = 130$) på baggrund af behandlingsforløb i antipsykotisk monoterapi i minimum 4 uger. Fastlæggelsen af markører sker i tilfælde af stærk overensstemmelse ($k > 0,6$) imellem to uafhængige fordelinger af samme kohorte; (1) opdelingen af patientkohorten i to grupper på baggrund af behandlingsudfaldet (forventet fordeling); (2) opdelingen af patientkohorten ved k-medoide clustering af udvalgte parametre (observeret fordeling). Trods gentagne undersøgelser med clusterfordelinger baseret på forskellige kliniske variable og selektionssnit af journalnotater opnåedes kun svag overensstemmelse ($k < 0,2$) mellem fordelingerne. Det konkluderes, at yderligere undersøgelser er nødvendige for at bestemme betydningen af informationer gemt i større databaser som EPJ-databasen for behandlingsudfald ved den antipsykotiske behandling.

Abstract

During admission to treatment at the Psychiatric Centre Sct. Hans, patients are monitored and information stored in medical records. Based on the hypothesis that parameters extracted from stored information in medical records will act as predictors of treatment outcomes, predictors were sought in a retrospective cohort study of 354 psychiatric patients with treatment outcomes of discharge ($n = 224$) and switch in medication ($n = 130$) following a minimum of 4 weeks of antipsychotic monotherapy. Predictors were assessed in case of substantial agreement ($k > 0,6$) between two independent distributions of the same cohort: (1) a distribution based on the treatment outcome (expected distribution); (2) a distribution based on k-medoids clustering of selected parameters (observed distribution). In spite of numerous cluster analyses based on different clinical variables and subsets of entries in medical records, only slight agreement ($k < 0,2$) was obtained between different distributions. As a conclusion, it is suggested that further research should be conducted to assess the impact on antipsychotic treatment outcomes of information contained within large repositories of medical records.

Forord

Sideløbende med udarbejdelsen af dette speciale har jeg bidraget til en monitorering af psykofarmakaordinationerne på Psykiatrisk Center Sct. Hans ud fra afdelingernes lægemiddelordinationer registreret i EPJ-systemet. Et tiltag der til mit kendskab kun har medført positive reaktioner fra klinikken. Nogle er sågar kommet med forslag til supplerende monitoreringer.
Årsagen til interessen skal findes i det hidtil ukendte overblik over psykofarmakaordinationerne, som i EPJ-systemet er registreret i overskuelige klumper af journaldata. En simpel forespørgsel på, hvem der en given dato modtog et givent præparat, vil således tage ugers gennemgang af afdelingernes journaldata at finde frem til. *Utroligt.* Hensigten med dette speciale er derfor at skabe fokus på det vidensaktiv, som de - ofte - uudforskede patientdata i EPJ-systemer repræsenterer. Samtidig er det en påmindelse til både planlæggere og leverandører om at stille krav til deres patientjournaler og administrationssystemer for at facilitere klinikkens adgang til disse data.

Indhold

Kapitel 1

Indledning

Der indsamles dagligt en større mængde data i landets elektroniske patientjournaler (EPJ). Alene på Psykiatrisk Center Sct. Hans og tidligere Sct. Hans Hospital blev der i løbet af 2007 oprettet 351.064 nye journalnotater[1] på baggrund af i alt 783 patienter tilknyttet centret[2]. Dette svarer til at hver patient i gennemsnit genererede 448 notater i 2007. Målt i antallet af notater for hele patientgruppen var registreringsfrekvensen i 2007 på mere end 960 notater om dagen. Et antal der hverken inkluderer ændringer af eksisterende notater eller medicinske registreringer (ordinationer, udleveringer og seponeringer).

Det er givetvis muligt for den trænede kliniker med et indgående kendskab til en lille patientgruppe og dennes patientjournaler at bevare overblikket men virkeligheden for den danske sundhedssektor er sandsynligvis en anden. I en artikel om EPJ i almen praksis i Ugeskrift for Læger skriver Ole Nordland: "i takt med, at informationsmængden vokser voldsomt ved at fremmede kolleger skriver i vores journaler, forsvinder vores overblik" [Nordland, 2001]. Selvom citatet er taget fra almen praksis så er antallet af forfattere på patientjournaler i primærsektoren ikke overraskende langt større. På Psykiatrisk Center Sct. Hans er det alle personalegrupper med patientkontakt (læger, sygeplejersker, psykologer, socialrådgivere, fysioterapeuter og ergoterapeuter) der agerer medforfattere i patientjournalerne. For det skal de[3].

I en nyligt publiceret undersøgelse, der sammenlignede papirsjournalerne med elektroniske patientjournaler i psykiatriske centre i staten Indiana, USA, var konklusionen den, at de elektroniske udgaver var bedre til at dokumentere den medicinske behandling [Tsai and Bond, 2008]. En undersøgelse der således slutter op om EPJ-Observatoriets bekendtgjorte nytteværdi ved EPJ om bedre dokumentation [Vingtoft et al., 2005].

En væsentlig udfordring for udnyttelsen af det elektroniske materiale i patientjournalsystemer som EPJ-systemet på Psykiatrisk Center Sct. Hans er imidlertid, at

[1] Se bilag A.1.1

[2] Se bilag A.1.2

[3] Sundhedsstyrelsen. Bekendtgørelse nr. 1373 af 12/12/2006 om lægers, tandlægers, kiropraktorers, jordemødres, tandplejeres, optikeres og kontaktlinseoptikeres patientjournaler (journalføring, opbevaring, videregivelse og overdragelse m.v.)[Smith, 2006]

sådanne systemer netop er fokuseret på dokumentation og ikke konstrueres med tanke på udnyttelse af informationen og den potentielt set vigtige opsparede viden.

1.1 Data Mining

Data mining er et informationsteknologisk begreb, der anvendes om de metoder, som knytter sig til uddragning af viden fra større database-systemer, hvor datamængden umuliggør brugen af gængse statistiske metoder. De metoder som data mining bringer i spil, har overordnet set til formål at identificere relationer og synliggøre strukturer i det pågældende datamateriale.
Tekst data mining er en specialisering af begrebet, der vedører dataindhold baseret på tekstdokumenter og samtidig også en specificering af de problemer, som metoderne skal tackle. Til forskel for nummeriske data rummer tekstelementer, der strækker sig fra strukturer med få ord til længere sammenhængende tekstklumper, en lang række fortolkningsmæssige udfordringer når det optræder i så enorme datasæt som eksempelvis elektroniske patientjournaler.

1.2 Medicinsk data mining

Anvendelse af data mining på klinisk materiale er en forholdsvis ny disciplin. Dette års udgave af tidskriftet *Yearbook of medical informatics* bragte en review-artikel af [Meystre *et al.*, 2008], der baseret på PubMed-listede artikler siden 1995 var i stand til at udvælge 174 veldokumenterede tekst mining undersøgelser af klinisk materiale. Men selvom [Meystre *et al.*, 2008] dokumenterede en stigende kvalitet i forhold til både metodisk tilgang og de anvendte analyseredskaber, fandt de stort set ingen anvendelser af teknologien uden for de enkelte forskningsprojekter.
En kort gennemgang af litteraturen afslører da også, at det hovedsageligt er *farmakovigilance* det vil sige medicinske risikoanalyser, hvor medicinsk data mining i dag har størst gennemslagskraft. En række review-artikler fra de sidste to år eksemplificerer dette fokus: [Bate *et al.*, 2008] præsenterede anvendelsen af medicinsk data mining på indrapporterede bivirkninger i WHO-regi; [Almenoff *et al.*, 2007] beskrev nye statistiske metoder i relation til eksisterende farmakovigilance-anvendelser; [Choi and Park, 2007] beskrev status i Sydkorea for indrapporteringer og muligheder for opbygningen af et nationalt farmakovigilance-system mens [Iskander *et al.*, 2006] kommenterede på situationen i USA med brugen af teknologien på et afrapporteringssystem for bivirkninger ved vaccinationer.
Den høje interesse for farmakovigilance i forhold til anvendelse af data mining på elektroniske patientjournaler kan have flere årsager, hvoraf kompleksiteten og begrænsede uddannelsesmæssige ressourcer formegentlig er kernefaktorer. Men selvom medicinsk data mining i patientadministrationssystemer endnu er på begynderstadiet, er potentialet åbenlyst. Til forskel fra farmakovigilance-analyser giver analyser af samtlige patientdata mulighed for at kortlægge sammenhænge, der kan gøres til genstand for såvel hypotese-generering som ny viden om både lidelser og behandlinger.

Hidtil ukendte sammenhænge mellem tidligere urelaterede lidelser kan således kaste nyt lys over særlige ætiologiske faktorer og give klinikere mulighed for i højere grad at praktisere evidensbaseret medicin.
[Mullins *et al.*, 2006] der havde sat sig for at undersøge potentialet for medicinsk data mining konstaterede på baggrund af deres resultater, at *usuperviserede* undersøgelser, repræsenteret ved data mining, genererede potentielt betydningsfulde strukturer i datasættet. Et datasæt der vel at mærke indeholdt informationer fra 667.000 patientforløb. At undersøgelserne foretaget af [Mullins *et al.*, 2006] var hypoteseløse vidner desuden om, at medicinsk data mining som automatiseret proces har potentialet til at blive anvendt i fremtidig klinisk forskning.

1.3 Relaterede undersøgelser

Specialets undersøgelser er baseret på en hypotesedreven metodisk tilgang, hvilket i stor udstrækning er gældende for de data mining undersøgelser der baserer sig på data fra patientadministrationssystemer, som jeg har fundet frem til ved generelle søgninger på Internettet samt søgninger på kliniske undersøgelser i PubMed. Disse *relaterede undersøgelser* har imidlertid i store træk alle haft til formål, at undersøge en bestemt type behandling eller en specifik lidelse på baggrund af enkeltvis registrerede hændelser. Dette står i modsætning til specialets undersøgelser, hvor formålet er at undersøge baggrunden for behandlingsudfald med udgangspunkt i både journalnotater fra patientkohortens behandlingsforløb[4] samt udvalgte kliniske variable, der foreligger ved behandlingsforløbets start[5].
En anden væsentlig forskel består i de valgte data mining metoder, idet de relaterede undersøgelser primært anvender associationsanalyser[6]; Analyser som i øvrigt inddrages ud fra nogle præfabrikerede værktøjer som eksempelvis *HealthMiner* eller *CliniMiner*, der blev anvendt af [Mullins *et al.*, 2006]. Dette begrænser naturligvis gennemskueligheden af den metodiske tilgang til data mining processen, der kræver et nærmere kendskab til de pågældende programmer samtidig med at rummeligheden i undersøgelsesdesignet begrænses af de krav som metodeapparatet fra de færdige programpakker stiller.
Den væsentligste forskel mellem de eksisterende undersøgelser og specialets ligger i den metodiske tilgang. Hvor de eksisterende undersøgelser fokuserer på identifikationen af hidtil ukendte sammenhænge i datasættet, har specialets undersøgelser til formål at åbenbare faktorer i det enorme datamateriale fra et patientadministrationssystem, der er prædikterende for væsentlige behandlingsudfald. En opgave der til mit kendskab ikke tidligere er forsøgt gennemført.

[4] Behandlingsforløb er i afsnit 2.4.5 defineret som perioden med samme kontinuerlige medicinske behandling.

[5] Formålet er uddybet i kapitel 2.

[6] Data mining metode der benyttes til at fremhæve samtidig tilstedeværelse af udvalgte variable i et datasæt med angivelse af et mål for den hyppighed hvormed disse samtidige forekomster observeres. Se afsnit 5.2.4.

Kapitel 2

Formål

Med dette speciale ønsker jeg at udpege markører i de elektroniske patientjournaler (EPJ) fra EPJ-systemet på Psykiatrisk Center Sct. Hans, der er udslagsgivende for vigtige behandlingsudfald i den farmakologiske antipsykotiske behandling[1].

2.1 Målsætning

Specialets undersøgelser bygger på en hypotesedreven strategi, hvor to uafhængige fordelinger af samme patientkohorte sammenlignes; (1) i første omgang opdeles patientkohorten i to grupper på baggrund af behandlingsudfaldet (forventet fordeling); (2) dernæst opdeles patientkohorten ved clustering af udvalgte parametre identificeret i EPJ (observeret fordeling).
Undersøgelserne gennemføres ved at sammenholde den forventede fordeling med forskellige observerede fordelinger på baggrund af de valgte parametre.
Markører for behandlingsudfald udgøres af de parametre, der ved opdeling af patientkohorten medfører en stærk overensstemmelse ($k > 0,6$)[2] mellem den observerede og den forventede fordeling.

2.2 Baggrundshypoteser

I dette afsnit præsenteres de tre overordnede hypoteser der ligger til grund for specialets undersøgelser:

- Journaldata indeholder parametre for behandlingsudfald
- Udskrivning og behandlingsskifte er to vigtige, komplementære og gensidigt ekskluderende udfald
- Parametre for behandlingsudfald udgøres af journalnotater og kliniske variable

[1] Udover den farmakologiske terapi (behandlingen med lægemidler) indgår også kognitiv terapi samt sociale og rehabiliterende interventioner i den antipsykotiske behandling [Fowler, 2000].

[2] Anvendelsen af Kappa-værdier som mål for det observerede udfalds overensstemmelse med den forventede clusterfordeling er gennemgået i afsnit 7.1 side 94.

2.2.1 Journaldata indeholder parametre for behandlingsudfald

Den elektroniske patientjournal (EPJ) har blandt andet til formål[3] at danne grundlag for behandlingen af patienten og dokumentere den udførte behandling. Det foranlediger hypotesen, at de journalførte oplysninger (journaldata) i EPJ indeholder parametre, der udgør mulige markører for udfaldet af den pågældende behandling. Det er derfor min hypotese, at sådanne parametre vil være mulige at erkende i journaldata indsamlet fra behandlingsforløb[4] førende til et af to vigtige, komplementære og gensidigt ekskluderende behandlingsudfald.

2.2.2 Udskrivning og behandlingsskifte er to vigtige, komplementære og gensidigt ekskluderende udfald

Til specialets undersøgelser er udvalgt behandlingsudfaldene *udskrivning* og *behandlingsskifte*, der udover at være gensidigt ekskluderende betragtes som komplementære på baggrund af følgende hypoteser; (1) at udskrivning af patienten som minimum er betinget af en stabilisering af den psykotiske lidelse, mens (2) behandlingsskifte med overgang til anden antipsykotisk behandling er begrundet i et eller flere af følgende; (2.1) fravær af antipsykotisk effekt, (2.2) uberettiget høj bivirkningsforekomst i forhold til behandlingens antipsykotiske effekt, (2.3) forventning om væsentlig forbedring af den antipsykotiske effekt ved behandlingsskifte, (2.4) svigtende komplians[5], (2.5) forværring af psykosen eller (2.6) patientens ønske om behandlingsskifte.
Disse behandlingsudfald er yderligere vigtige, idet udskrivning i forlængelse af det foregående kan betragtes som et *positivt* udfald af behandlingen, mens behandlingsskifte er kendetegnet ved fraværet af den positive effekt og således er at betragte som et *negativt* behandlingsudfald.

2.2.3 Parametre for behandlingsudfald udgøres af journalnotater og kliniske variable

På baggrund af de tilgængelige kliniske data registreret i EPJ-systemet er det sluttelige min hypotese, at de parametre der udgør markører for behandlingsudfaldene er indeholdt i henholdsvis (1) *journalnotater*[6] registreret i løbet af behandlingen frem til behandlingsudfaldet samt i (2) tilstedeværende *kliniske variable* ved behandlingens påbegyndelse. Dette er begrundet i følgende hypoteser; (1) journalnotater registreres kontinuerligt i løbet af behandlingen og er dermed beskrivende for både den fænotypiske præsentation (det vil sige det observerede sygdomsudtryk) hos patienten, de kliniske interventioner (det vil sige den valgte behandling der søger at mildne symtomerne) samt den konkluderende fortolkning af behandlingseffekten. Dermed vil et

[3] Se afsnit 4.2.1.

[4] Behandlingsforløb er i afsnit 2.4.5 defineret som perioden med samme kontinuerlige medicinske behandling.

[5] Komplians er af [Awad and Voruganti, 2004] beskrevet som graden af patientens efterlevelse af de kliniske retningslinier herunder den medicinske behandling.

[6] Se afsnit 4.4.

givent selektionssnit[7] af journalnotater kunne benyttes til at udtrykke det rationale, der knytter sig til den igangværende behandling og vil i forlængelse heraf indeholde markører for det pågældende behandlingsudfald; (2) de kliniske variable køn, alder, afdelingstilknytning og diagnose er alle af betydning for den valgte intervention: (2.1) Køn er en betydende parameter i forhold til en række interventioner herunder også andre for behandlingen tilstedeværende terapiformer end den farmakologiske. På den farmakologiske side er der for kvinders vedkommende eksempelvis påvist en højere grad af bivirkninger ved behandling med atypiske antipsykotika[8] [Aichhorn *et al.*, 2000]. (2.2) Alder er en anden parameter, der ligesom køn er påvist at have betydning for bivirkningsforekomsten ved den antipsykotiske behandling [Newcomer, 2004]. (2.3) Afdelingstilknytning er af væsentlig betydning for den valgte kliniske intervention[9] og dermed også behandlingsudfaldet. (2.4) Diagnosen er baseret på symptombilledet[10] og er dermed en betydende parameter for valg af intervention og således også udfaldet af den antipsykotiske behandling.
Udsagnet af ovenstående hypoteser er skitseret på figur 2.1 og danner baggrund for opstillingen af problemformuleringen i afsnit 2.3.

2.3 Problemformulering

Hvilke markører kan identificeres i databasen for elektroniske patientjournaler (EPJ) på Psykiatrisk Center Sct. Hans som prædikterende for behandlingsudfaldene udskrivning og behandlingsskifte ved psykiatriske diagnoser, hvor behandling med antipsykotika er indiceret?

2.3.1 Præcisering af problemformuleringen

Besvarelsen af ovennævnte problemformulering fordrer en præcisering af dens metodiske implikationer:

1. Markører er de parametre, i form af enten kliniske variable og/eller et selektionssnit af journalnotater, der er udslagsgivende for behandlingsudfaldet.

2. Identifikationen af parametre i journaldata for behandlingsudfald kræver en fastlæggelse af den periodemæssig afgrænsning af behandlingsforløb der forventes at rumme markører for udskrivning og behandlingsskifte.

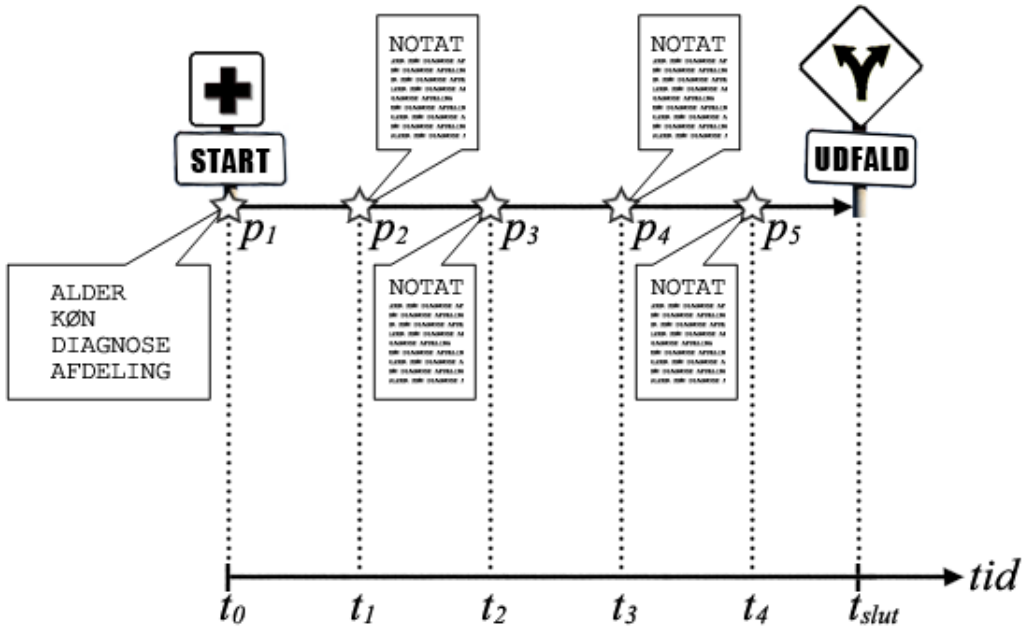
3. For undersøgelsen af parametrenes betydning for behandlingsudfaldet kræves fastlæggelsen af en sammenlignelig patientkohorte hvis behandlingsforløb, baseret på journaldata i EPJ-systemet;

[7] Med selektionssnit menes et specifikt udsnit af journalnotaterne for det pågældende behandlingsforløb. Dette udsnit kan enten udgøre hele mængden eller blot en delmængde af forløbets journalnotater.

[8] For klassifikation af antipsykotika se afsnit 3.3.2.

[9] Afdelingerne og deres arbejdsområder er gennemgået i afsnit 4.1.2.

[10] ICD-10 diagnosticeringen er gennemgået i afsnit 3.1.2.



Figur 2.1: Et skitseret behandlingsforløb (øverste vandrette pil) baseret på registreringer i EPJ-systemet. Behandlingsforløbet fører til et af to mulige udfald ved forløbets ophør (tidspunktet på figurens tidsakse er angivet som t_{slut}). Specialets undersøgelser er baseret på den hypotese, at parametre (p_x) i de tilstedeværende journaldata fra behandlingsforløbets start (t_0) til dets ophør (t_{slut}) fungerer som markører for udfaldet af det pågældende behandlingsforløb. Disse parametre udgøres henholdsvis af kliniske variable (p_1), der er tilstede ved behandlingens start (t_0) og/eller af et selektionssnit af behandlingsforløbets journalnotater, repræsenteret ved parametrene p_2-p_5 til tiderne t_1-t_4.

(a) fører til enten udskrivning eller behandlingsskifte, således at kohorten kan opdeles i to distinkte grupper på baggrund af behandlingsudfaldet benævnt den forventede fordeling.

(b) indeholder en eller flere parametre, der benyttet som inputdata ved clustering fører til opdelingen af patientkohorten i to distinkte grupper benævnt den observerede fordeling.

4. Overensstemmelse mellem (a) den forventede fordeling og (b) den observerede fordeling udtrykkes som Kappa-værdien for denne overensstemmelse.

5. Markører er de parametre, der fører til en stærk overensstemmelse ($k > 0,6$) mellem den forventede fordeling og den observerede fordeling.

2.4 Afgrænsning

På baggrund af problemformuleringen fastlægges i dette afsnit afgræsningen for specialets undersøgelser, der knytter sig til det valgte empiriske materiale. Afgrænsingen har til formål (1) at sikre så stor en patientkohorte til specialets undersøgelser som muligt, (2) at sikre så veldokumenteret en patientkohorte som muligt, samt (3) at sikre så sammenlignelige behandlingsforløb som muligt.

2.4.1 Rekrutteringsperiode

Idet specialet gennemføres på Forskningsinstituttet på Psykiatrisk Center Sct. Hans indenfor rammerne af centrets kvalitetsudvikling, udgøres det tilgængelige datamateriale af samtlige registreringer i EPJ-systemet siden dets introduktion i 1997[11]. Selvom EPJ reelt var i fuldt brug fra december 2002 [Kjeldsen, 2002] på alle centrets afdelinger har overgangen til de elektroniske registreringer ikke været helt problemfri. Der er langt større konsistens i de nyligst registrerede data, ligesom specificering af eksempelvis dosisstørrelse først indtræder på datasættet generelt fra medio 2005.
For at sikre så stor en patientgruppe som mulig, samtidig med at den størst mulige konsistens i de journalførte oplysninger tilstræbes, har jeg valgt perioden fra 1. april 2003 til 1. april 2008 som rekrutteringsperiode.

2.4.2 Diagnoser

Indikationer for behandling med antipsykotiska er primært skizofreni og de skizofrenirelaterede lidelser [WHO, 1993]. En række psykiatriske lidelser er imidlertid også karakteriseret ved tilstedeværelsen af psykotiske symptomer, herunder både affektive og misbrugsrelaterede lidelser [Mauri *et al.*, 2005]. Indenfor rammerne af ICD-10 diagnosesystemet, som fastsat af [WHO, 1993], har jeg derfor udvalgt F 1-3 diagnoserne[12] som diagnosekrav til patientkohorten.

[11] Implementeringen af EPJ-systemet skete i første omgang som pilotprojekt.
[12] Diagnoserne er uddybet i afsnit 3.1.2 side 16.

2.4.3 Antipsykotika

Psykosernes symptombehandling sker med et eller flere antipsykotika. For undersøgelsen inkluderes alene præparater, der på ordinationstidspunktet er klassificeret under lægemiddelkatalogets[13] terapeutiske gruppe af antipsykotika uanset administrationsmetode[14]. De anvendte præparatbetegnelser er angivet i tabel 2.1.
Behovsmedicinering[15] (*pro necessitate* medicin) ekskluderes dog fra præparatselektionen ud fra den hypotese, at det er de præparater, som gives fast, der er udslagsgivende for den generelle medicinske behandlingseffekt.

2.4.4 Antipsykotisk monoterapi

For de til undersøgelsen valgte behandlingsforløb, skal der være tale om antipsykotisk monoterapi. Monoterapi defineres som den medikamentelle behandling, hvori der kun optræder ét præparat. I specialet har denne definition dog kun betydning for behandlingen med antipsykotika. Det skal dog bemærkes at behovsmedicinering samt præparater der ikke er klassificeret som antipsykotika ikke medregnes ved bestemmelse af monoterapiforløb.
Som betegnelse for flere samtidige præparatordinationer benyttes *polyfarmaci*. Monoterapi og polyfarmaci er eksemplificeret i figur 2.2 der illustrerer et ordinationsskema for en patient med tre ordinationer. På figuren er der på 3. og 7. dagen tale om monoterapi med henholdsvis præparat B og præparat C, mens der på de øvrige dage er tale om polyfarmaci med to (2. og 4. dagen) eller tre præparater (1., 5. og 6. dagen).

Figur 2.2: Et skitseret ordinationsskema for en patient, der i løbet af syv døgn modtager præparaterne A, B og C. Der er på 3. og 7. dagen tale om monoterapi med henholdsvis præparat B og præparat C, mens der de øvrige dage er tale om polyfarmaci med to (2. og 4. dagen) eller tre præparater (1., 5. og 6. dagen).

Det skal bemærkes at afgrænsningen monoterapi alene knytter sig til de for undersøgelserne valgte behandlingsforløb. Det er således uden betydning om patienten efter udfaldet af det valgte behandlingsforløb overgår til antipsykotisk polyfarmaci.

[13] Medicin.dk afløste fra oktober 2006 det medicinske opslagsværk Lægemiddelkataloget®, der omfatter alle markedsførte humane præparater. Medicin.dk er derfor anvendt som udgangspunkt for identifikation af antipsykotika.

[14] Se afsnit 3.3.1 side 24.

[15] Behovsmedicinering er en slags *stand by* ordinationer, som sygeplejersken kan benytte i tilfælde af akut behov. Det kan eksempelvis være et antipsykotikum som anvendes på grund af dets beroligende effekt til dæmpe udadreagerende eller aggressiv adfærd.

	Administrationsmetode	
	Konventionel	Depot
Typiske antipsykotika		
Fluanxol	✓	✓
Siqualone		✓
Serenase	✓	✓
Orap	✓	
Neulactil	✓	
Trilafon		✓
Stemetil	✓	
Cisordinol	✓	✓
Truxal	✓	
Nozinan	✓	
Buronil	✓	
Dipiperon	✓	
Dogmatil	✓	
Atypiske antipsykotika		
Solian	✓	
Abilify	✓	
Clozapin	✓	
Leponex	✓	
Zyprexa	✓	
Invega	✓	
Seroquel	✓	
Risperdal	✓	✓
Serdolect	✓	
Zeldox	✓	
Semap		✓

Tabel 2.1: Præparatnavne for antipsykotika med angivelse af præparaternes administrationsmetode frit efter [Pedersen *et al.*, 2008]. Den konventionelle administrationsform henviser til kapsler og tabletter med antipsykotisk virkning i højest 48 timer mens depot henviser til tablet eller intramuskulær injektion med antipsykotisk virkning i 7-14 døgn.

Årsagen til at polyfarmaci imidlertid ekskluderes fra de valgte behandlingsforløb er den, at evidensen for behandlingens virkning er meget ringe [Tranulis *et al.*, 2008] ligesom kombinationsbehandlinger øger risikoen for blandt andet farmakokinetiske interaktioner[16] [Miller and Craig, 2002]. Ved at ekskludere polyfaramci styrkes evidensen for de valgte behandlingsforløb ligesom sammenligneligheden mellem patientkohortens behandlingsforløb øges.

Problematikken som antipsykotisk polyfarmaci rejser vil blive yderligere gennemgået i afsnit 3.4.2.

2.4.5 Behandlingsforløb

Behandlingsforløb defineres som perioden med samme kontinuerlige medicinske behandling. For en patient, der i en periode modtager ét præparat og efterfølgende skifter til et andet præparat, er der dermed tale om to forskellige behandlingsforløb adskilt af et behandlingsskifte. Dette er eksemplificeret på figur 2.3 (1). Havde samme patient i stedet fået ordineret et supplerende præparat, ville der således være tale om et behandlingsskifte hvor patienten overgår fra monoterapi til polyfarmaci. Dette er eksmplificeret på figur 2.3 (2).

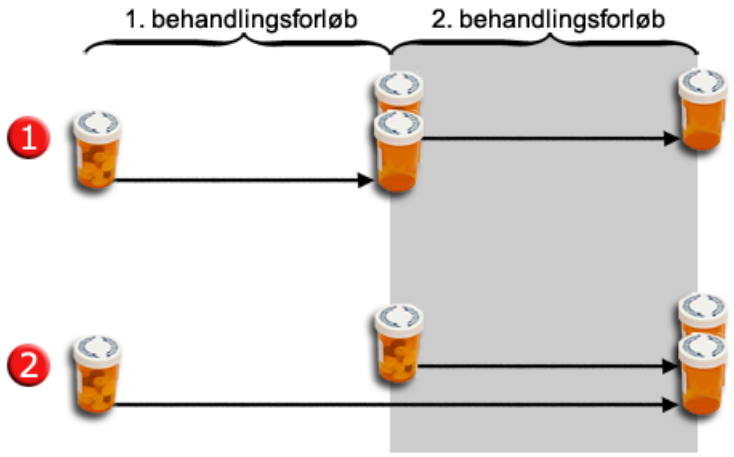
Ophør af et behandlingsforløb angives som et *behandlingsudfald*. Ved førnævnte eksempler er der i begge tilfælde tale om, at første behandlingsforløb afsluttes med udfaldet behandlingsskifte. Var patienten i stedet for skiftet blevet udskrevet, ville behandlingsforløbets udfald således være karakteriseret ved en udskrivning.

2.4.6 Behandlingsforløbets varighed

For at kunne betragte behandlingsforløbet som udslagsgivende for det pågældende behandlingsudfald, er det nødvendigt at fastlægge en minimumsvarighed for dette forløb. Idet varigheden af behandlingsforløbet er bestemt ved det ordinerede antipsykotikum, stilles således krav om, at effekten af det pågældende præparat som minimum skal være indtruffet ved tidspunktet for behandlingsforløbets udfald. Idet behandlingseffekten af det pågældende antipsykotikum både er præparat- og patientafhængig, har jeg valgt at adspørge overlæge Klaus Damgaard Jakobsen (KDJ), Psykiatrisk Center Hvidovre, om at tage stilling til den forventede generelle minimumsperiode for erkendelse af et antipsykotisk-respons på behandlingen. En sådan minimumsperiode har KDJ anbefalet fastsættes til 4 ugers medicinsk behandling.

KDJs anbefaling støttes af de generelle målepunkter for vurdering af den antipsykotiske behandling som finder anvendelse i litteraturen generelt [Chouinard *et al.*, 1978], [Manchanda and Hirsch, 1986], [Hill *et al.*, 1992], [Correll *et al.*, 2003] og [Emsley *et al.*, 2006]. Jeg har derfor fastlagt minimumsvarigheden af behandlingsforløbet til 4 uger.

[16] Det vil sige at de ordinerede præparater påvirker hinandens både ønskede og uønskede effekter.



Figur 2.3: Behandlingsforløbet er defineret som den samme kontinuerlige medicinske behandling. Figuren præsenterer to eksempler på behandlingsskifte; (1) opdeling baseret på to forskellige ordinationer, og (2) opdeling baseret på tilføjelsen af yderligere en ordinationen. Der er således i begge tilfælde tale om, at behandlingsforløbene er adskilt af et behandlingsskifte.

2.4.7 Behandlingsudfald

Som angivet i problemformuleringen[17] har jeg valgt udfaldene behandlingsskifte og udskrivning. Disse behandlingsudfald er gensidigt ekskluderende og repræsenterer to diametralt modsatte kliniske udfald, hvilket dermed styrker min forventning til undersøgelsens fordelinger. For at mindske bias[18] forbundet med den periodemæssige placering af behandlingsudfaldets optræden er det alene behandlingsforløb fra patientens første indlæggelse registreret i EPJ-systemet, som medtages. For behandlingsforløbet der leder frem til udfaldet „behandlingsskifte" skal det endvidere gælde, at der er tale om indlæggelsens første behandlingsforløb. Sammenhængen mellem indlæggelsens behandlingsforløb og udfaldene er illustreret på figur 2.4

2.4.8 Indlæggelser

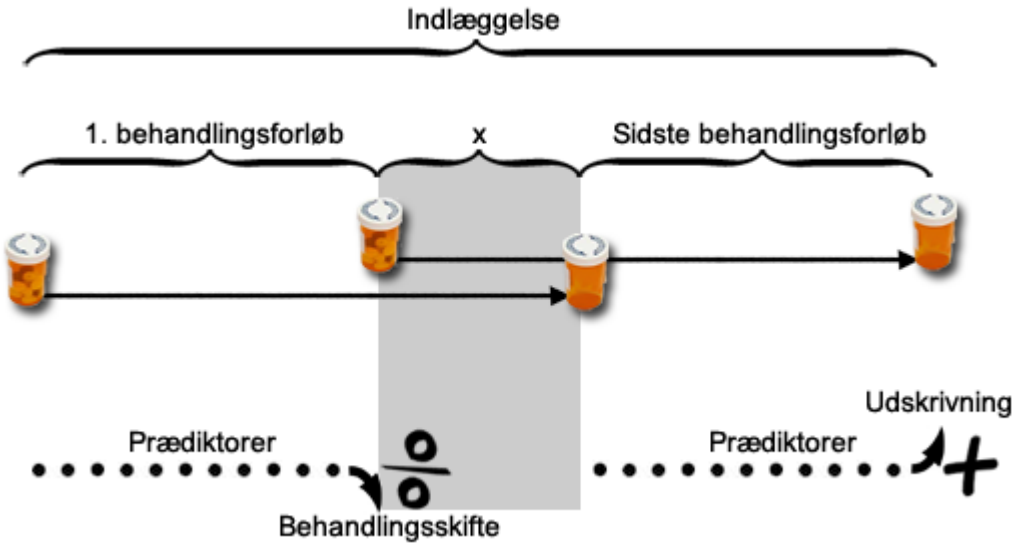
Kun første indlæggelse for indlæggelsesforløb på sengeafsnit medtages i undersøgelsen for at mindske bias i form af eventuelle erfaringer fra tidligere indlæggelser hos både patient og behandler. Alene afsluttede indlæggelser medtages i undersøgelsen, idet alle behandlingsudfald for undersøgelsesgruppen skal være kendte ved forsøgets start.

2.5 Inklusionskriterier for patientkohorten

Afgrænsningen fører til opstillingen af følgende inklusionskriterier:

[17] Se afsnit 2.3.

[18] På baggrund af bias definitionen præsenteret af [Greenfield *et al.*, 1996] defineret bias i relation til specialets undersøgelser, som den forudindtagede personlige præference, der påvirker tilgangen til den pågældende behandling.



Figur 2.4: Indlæggelse opdelt i perioder af behandlingsforløb, hvor første behandlingsforløb fører frem til et behandlingsskifte, og det sidste og afsluttende behandlingsforløb fører frem til patientens udskrivning. At det mellemliggende (grå) område ekskluderes skyldes både at behandlingsforløbet foregår i polyfarmaci samt at det ikke fører frem til et udfald i specialets undersøgelser, der alene inddrager behandlingsskifte fra første behandlingsforløb samt behandlingsforløb førende til udskrivelse.

- Patientgrundlaget udgøres af patienter med afsluttede indlæggelser registreret på sengeafsnit på Psykiatrisk Center Sct. Hans i perioden fra 1. april 2003 til 1. april 2008
- Patienter skal have minimum en af følgende diagnoser: F1, F2 eller F3
- Patienter bidrager med journalførte oplysninger fra behandlingsforløb registreret under patientens *første* indlæggelse
- Behandlingsforløb skal udgøres af minimum 4 ugers antipsykotisk monoterapi
- Behandlingsforløb skal føre til enten behandlingsskifte eller udskrivning; for behandlingsforløb førende til behandlingsskifte skal der være tale om indlæggelsens første behandlingsforløb
- Patienten bidrager med kun *et* behandlingsforløb. Hvis patientens indlæggelse kan bidrage med flere behandlingsforløb til undersøgelsen selekteres alene det første behandlingsforløb

2.6 Empirisk grundlag

Det empiriske grundlag for specialets undersøgelser udgøres af samtlige registreringer i EPJ-systemet. For perioden fra 1. april 2003 til 1. april 2008 udgøres registrerings-

grundlaget således af 1987 patienter[19] fordelt på 667 kvinder og 1320 mænd[20].
De anvendte inklusionskriterier medfører dog et væsentligt selektionssnit, hvilket betyder at det reelle empiriske grundlag ikke kan bestemmes direkte på baggrund af det tilgængelige datamateriale i EPJ-systemet. Jeg har derfor valgt at gennemføre en præliminær dataundersøgelse, både for at fastlægge det empiriske grundlag og til brug for specialets senere undersøgelser. Denne indledende dataundersøgelse betegnes specialets pilotforsøg[21] og på baggrund af denne fastlægges empirien til at bestå af journalførte data for 130 behandlingsforløb førende til et behandlingsskifte og 224 behandlingsforløb førende til udskrivning. Disse i alt 354 behandlingsforløb tilfredsstiller inklusionskriterierne som fastlagt ovenfor. Forsøgets patientkohorte benyttes som synonymbetegnelse for de identificerede behandlingsforløb, idet hvert forløb alene vedrører én patient.
Væsentligste begrænsning for identifikationen af behandlingsforløb var fraværet af dosisstørrelser på ordinationer for store dele af datasættet. Dosisstørrelser indgår derfor ikke i specialets undersøgelser.

2.7 Overordnede udfordringer

Opstillingen af et hypotesedrevet forsøg rettet mod registreringerne i EPJ-systemet indeholder en række datalogiske og medicinalbiologiske udfordringer.
Medicinalbiologisk koncentrerer udfordringerne sig primært mod forståelsen af den kliniske adfærd, kendskab til behandlingsmæssige problematikker og benyttelsen af denne baggrundsviden til at fastlægge forsøgsparametre i det tilgængelige datamateriale, at bestemme forventninger til forsøgsudfald ud fra de valgte parametre og fortolkningen af de observerede forsøgsudfald.
Datalogisk knytter disse udfordringer sig særligt til fortolkningen af de tilgængelige data i EPJ-systemet og strukturering af disse data i relation til de valgte parametre samt valg af metoderedskaber til brug for dataundersøgelsen af de forskellige forsøgsparametre.
Idet specialet kombinerer fagområderne datalogi og medicinalbiologi er en anden væsentlig udfordring i dette brydningsfelt det formidlingsmæssige aspekt ved specialets undersøgelser.

2.8 Specialets opbygning

Specialet dokumenterer undersøgelsen af to grupper af behandlingsforløb og parametre fastlagt herfra på baggrund af data i EPJ-systemet.
Målgruppen for specialets undersøgelser er sundheds- og it-professionelle med interesse for videnuddragning baseret på data mining i patientadministrationssystemer.
Specialerapporten er inddelt i 9 kapitler og indeholder desuden bilagene A, B og C.

[19] Se bilag A.1.3
[20] Se bilag A.1.4
[21] Se metodens afsnit 6.3

Formål. Kapitel 1 og 2 præsenterer baggrunden for specialet med fastlæggelse af hypotese, undersøgelsesmål og problemformulering.

Teoretisk baggrund. Kapitel 3 til 5 præsenterer den teoretiske baggrund for tilrettelæggelsen af specialets undersøgelser med beskrivelse af den antipsykotiske behandling, EPJ-systemet og metoder knyttet til data mining.

Undersøgelser. Kapitel 6 og 7 beskriver den metodiske tilgang til specialets undersøgelser samt resultaterne heraf.

Diskussion. Kapitel 8 indeholder diskussionen af undersøgelsesudfaldene og en diskussion af den valgte tilgang til undersøgelsen.

Konklusion. Kapitel 9 indeholder konklusionen på specialets undersøgelser samt perspektivering.

Bilag. Bilag A dokumenterer datafremstillingerne i specialet mens bilag B og C dokumenterer specialets undersøgelser. Bilag B præsenterer dataanalyserne og bilag C analysernes resultater i form af clusterfordelinger.

Kapitel 3

Behandling med antipsykotiske lægemidler

Kapitlet har til formål at introducere til anvendelsen af antipsykotiske præparater, der er kendetegnet ved de behandlingsforløb, som blev præsenteret i afsnit 2.4. Der vil i kapitlet blive lagt vægt på brugen af kliniske termer, idet de parametre fra journalnotater i EPJ-systemet som specialets undersøgelser har til hensigt at inddrage, i høj grad gør brug af en sådan terminologi. Disse kliniske termer vil dog blive forklaret undervejs.

3.1 Indikation for anvendelse af antipsykotiske lægemidler

I dette afsnit præsenteres de diagnoser hvor behandling med antipsykotiske lægemidler er indiceret.

3.1.1 Begrebsdefinition

Antipsykotika benyttes i daglig tale som betegnelse for de lægemidler, der anvendes til antipsykotisk behandling. Selvom de psykiatriske behandlingsformer inkluderer andet end medicinske behandlinger, er det netop behandlinger med antipsykotika, der er udgangspunkt for bestemmelsen af de behandlingsforløb, der er i fokus i specialet. Den farmakologiske behandling med antipsykotika vil derfor blive omtalt som *antipsykotisk behandling*.

3.1.2 Diagnoser

Behandling med antipsykotika er en opgave, som varetages af speciallæger i psykiatri. På baggrund af en grundig udredning af de tilstedeværende symptomer diagnosticeres patientens psykotiske tilstand efter kriterierne i ICD-10. ICD-10 er en international standard fastsat i WHO-regi for klassifikation af helbredsrelaterede lidelser

[WHO, 1993]. Lidelserne er grupperet i 22 kapitler med tilhørende hovedgrupper og undergrupper. Hver lidelse er klassificeret ved hjælp af en diagnosekode, der angiver lidelsens placering i ICD-10 systemet. Diagnosekoden består af en bogstavkode og derpå en talrække bestående af op til tre tal. Jo flere tal der benyttes i angivelsen, jo højere er specificiteten af den pågældende lidelse. For de psykiske lidelser gælder, at de alle er grupperet under bogstavkoden F.

For at sikre så korrekt en klassificering som muligt, omfatter ICD-10 systemet desuden bestemmelser om kriterier for korrekt diagnosticering. Idet en række symptomer herunder også de psykotiske symptomer kan forekomme på tværs af diagnoserne, er der i visse tilfælde knyttet et diagnosticeringskrav om en indledende udelukkelse af relaterede lidelser.

Behandling med antipsykotika er primært indiceret ved psykotiske symptomer [Mauri *et al.*, 2005], hvilket indenfor ICD-10 klassifikationen hovedsageligt omfatter de skizofreni-relaterede lidelser, der har diagnosekoden F2. Idet psykotiske episoder også rammer patienter uden for dette spektrum, er diagnosekravet for patientkohorten i specialet udvidet til også at indbefatte de stofrelaterede psykiske lidelser (F1) og de affektive sindslidelser (F3).

3.1.3 F1 (Stofrelaterede psykiske lidelser)

Denne gruppe betegner psykiatriske lidelser forårsaget af stof- eller alkoholmisbrug og er listet op i tabel 3.1 [WHO, 1993].

Diagnosekode	Lidelse
F 10	Psykiatrisk lidelse som følge af alkohol
F 11	Psykiatrisk lidelse som følge af opioidemisbrug (narkotika)
F 12	Psykiatrisk lidelse som følge af cannabinoidemisbrug (hash)
F 13	Psykiatrisk lidelse som følge af beroligende stoffer eller sovemedicin
F 14	Psykiatrisk lidelse som følge af kokainmisbrug
F 15	Psykiatrisk lidelse som følge af andre stimulanter som koffein
F 16	Psykiatrisk lidelse som følge af hallucinogener (LSD og svampe)
F 17	Psykiatrisk lidelse som følge af tobak
F 18	Psykiatrisk lidelse som følge af opløsningsmidler
F 19	Psykiatrisk lidelse som følge af blandet stofmisbrug

Tabel 3.1: F10-19 diagnoserne [WHO, 1993].

3.1.4 F2 (De skizofreni-relaterede lidelser)

Denne gruppe udgør de væsentligste lidelser med indikation for antipsykotika og indbefatter alle skizofrenirelaterede lidelser. Gruppen er listet op i tabel 3.2 [WHO, 1993].

Diagnosekode	Lidelse
F 20	Skizofreni
F 21	Skizotypisk sindslidelse
F 22	Paranoide psykoser
F 23	Akutte og forbigående psykoser
F 24	Induceret psykose
F 25	Skizoaffektive psykoser
F 28	Anden non-organisk psykose
F 29	Uspecificeret non-organisk psykose

Tabel 3.2: F20-29 diagnoserne [WHO, 1993].

3.1.5 F3 (De affektive sindslidelser)

Denne gruppe udgøres af de affektive lidelser og er listet op i tabel 3.3 [WHO, 1993].

Diagnosekode	Lidelse
F 30	Manisk enkeltepisode
F 31	Bipolar affektiv sindslidelse
F 32	Depressiv enkeltepisode
F 33	Tilbagevendende periodisk depression
F 34	Vedvarende kroniske affektive lidelser
F 38	Andre affektive lidelser
F 39	Uspecificeret affektiv lidelse

Tabel 3.3: F30-39 diagnoserne [WHO, 1993].

3.1.6 Psykoser

Det, at være psykotisk, henviser til tabet af realitetssansen og til at den medførte ændrede virkelighedsopfattelse ikke kan korrigeres.
De psykotiske symptomer kan opstå pludseligt eller bryde gradvist frem over tid. Symptomerne er vidt forskellige fra patient til patient og varierer ofte også for patienter med gentagende psykotiske episoder. Symptomerne opdeles overordnet set i positive, negative og kognitive symptomer [Lambert and Castle, 2003]. Svært psykotiske patienter med forekomst af flere sådanne symptomer vil som følge af diagnosticerings kriterierne i ICD-10 systemet diagnosticeres som skizofrene [WHO, 1993]. I denne sammenhæng skal det derfor pointeres, at skizofreni og de skizofreni-relaterede lidelser hører til blandt de absolut alvorligste psykiatriske tilstande.

Positive symptomer

At symptomerne beskrives som positive skyldes at de er *kommet til* i forbindelse med sygdommens progression. Disse symptomer udgøres af vrangforestillinger (urealistiske

overbevisninger) og hallucinationer (det at sanse noget ikke-eksisterende) [Lambert and Castle, 2003].

Negative symptomer

De negative symptomer betegner de symptomer der kommer til udtryk som en *reduktion* af patientens funktionsevne. Symptomerne angivet af [Lambert and Castle, 2003] inkluderer:

- Ligegyldig emotionel tilstand (fravær af følelsesmæssige reaktioner)
- Anhedoni (fravær af evne til at føle nydelse)
- Asocialitet (social tilbagetrækning)
- Apati (ligegyldighed overfor alt)
- Alogia (talebesvær)
- Opmærksomhedsforstyrrelse (fravær af koncentrationsevne)

Kognitive symptomer

Disse symptomer betegnes også som neurodegenerative og kommer til udtryk ved en påvirkning af patientens *tænkning*. Ligesom de negative symptomer er der tale om et direkte *hæmmende* sygdomsudtryk og symptomer herfra klassificeres derfor også ofte som negative. Den væsentligste forskel er dog tidspunktet for symptomets indtræden, idet de kognitive symptomer siges at være tilstede ved sygdomsudbrudet mens de negative symptomer fremkommer i løbet af sygdommen [Lambert and Castle, 2003]. Disse symptomer inkluderer:

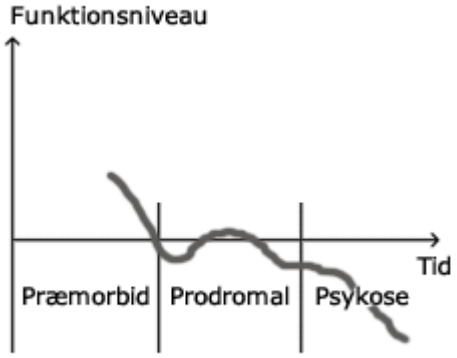
- Dissorganiseret adfærd (besvær med at løse simple problemstillinger)
- Opmærksomhedsforstyrrelse (fravær af koncentrationsevne)
- Hukommelsessvigt (genkendelses- og forståelsesvanskeligheder)
- Dissorganiseret tale (uforståelig tale- og ordbrug)

Non-psykotiske ledsagesymptomer

Psykoser ledsages ofte af non-psykotiske symptomer såsom selvmordstanker og selvmordspræget adfærd, affektive symptomer (bizarre humørsvingninger) og disorganiseret adfærd [Pantelis and Lambert, 2003].

3.2 Det skizofrene spektrum

Sygdomsforløbet for patienter i skizofrenispektret er beskrevet ud fra faseinddelingen i en præmorbid, en prodromal og en psykotisk sygdomsfase. Faseopdelingen er baseret på tilstedeværelsen af symptomer, idet de skizofrene lidelser er karakteriseret ved et progredierende sygdomsforløb, der munder ud i en fulminant psykotisk lidelse [Lewis and Lieberman, 2000].



Figur 3.1: Forløbet af det skizofrene sygdomsforløb med faldende funktionsevne som funktion af tiden. Selv før indtræden i den psykotiskesygdomsfase er der tale om et progredierende sygdomsforløb. I den præmorbide fase fremkommer de første sygdomstegn som gradvis forværres mens den prodromale fase kendetegnes ved fremkomsten af de egentlige skizofrene symptomer. Studier viser at, den største progression i sygdommen sker indenfor fem år af de første egentlige psykotiske symptomer [Sanger *et al.*, 1999].

Figur 3.1 illustrerer et skitseret sygdomsforløb med faldende funktionsevne over tid med angivelse af faseinddeling for skizofreni. Funktionsevnen skal her forstås er en fælles betegnelse for patientens mentale og sociale evner, som sætter patienten i stand til at opretholde en normal tilværelse. På figuren er den faldende funktionsevne ikke angivet som et lineært forløb. Dette skyldes, at det skizofrene sygdomsforløb er karakteriseret ved fremkomsten og forsvinden af psykotiske symptomer sideløbende med en række tiltagende sproglige-, kognitive- og adfærdsmæssige forstyrrelser [Frangou and Byrne, 2000].

Den præmorbide fase

Den præmorbide fase er karakteret ved et let nedsat funktionsniveau, der giver sig udslag i forskellige sociale, motoriske og/eller kognitive forstyrrelser [Lewis and Lieberman, 2000]. Et andet væsentligt karakteristika ved denne fase er fraværet af de egentlige skizofrene symptomer, hvilket har som konsekvens, at fasen reelt først kan erkendes retrospektivt.

Den prodromale fase

Den prodromale fase er karakteriseret ved en forværring af de præmorbide symptomer og/eller pludselige adfærdsmæssige forandringer hos patienten [Lewis and Lieberman, 2000]. Tvangstanker og tankeforstyrrelser er mere fremtrædende og patientens sproglige færdigheder påvirkes: „Ordenes betydning glider fra den gængse til en for patienten privat mening og nye sære ord, som kun patienter kender, kan dukke op..." [Stenstrøm *et al.*, 2008].

Den psykotiske sygdomsfase

Den psykotiske sygdomsfase er karakteriseret ved tilstedeværelsen af de egentlige skizofrene symptomer; positive, negative og kognitive [Lewis and Lieberman, 2000]. Symptombilledet er variende og den følgende gennemgang af lidelser indenfor det skizofrene spektrum suppleres derfor med anonymiserede uddrag af journalnotater fra patienter[1] på Psykiatrisk Center Sct. Hans.

3.2.1 Skizofreni (F20)

Skizofreni er en samlebetegnelse for lidelser med et symptombillede, der blandt andet manifesterer sig ved de såkaldte positive og negative symptomer samt kognitive forstyrrelser. De positive symptomer omfatter vrangforestillinger og hallucinationer, imens de negative symptomer omfatter de direkte sociallivsfortrængende symptomer som initiativløshed og isolation [Lewis and Lieberman, 2000].
Den psykotiske sygdomsfase indtræder almindeligvis først i voksenalderen, hvor patienter ofte debuterer fra slutningen af 20'erne til starten af 30'erne. Skizofreni optræder dog stort set i alle aldre [Lewis and Lieberman, 2000]. Selvom ætiologien bag skizofreni endnu ikke er fuldt klarlagt, viser nyere forskning at genetiske mutationer er involveret i udviklingen af skizofreni [Stefansson *et al.*, 2008].

Diagnostisering baseres på tilstedeværelsen af enten [WHO, 1993]:
Et eller flere førsterangs-symptomer:

- Tankepåvirkningsoplevelser (f.eks. tankeopslugning)
- Styringsoplevelser (f.eks. påførte viljeimpulser)
- Hørelsesallucinationer (f.eks. kommenterende stemmer)
- Andre hallucinationer (f.eks. vejrstyringsevner)

Og/eller to eller flere vrangforestillinger:

- Vedvarende hallucinationer med vrangforestillinger uden affektivt indhold
- Sproglige tankeforstyrrelser

[1] Der er for alle patienter tale om behandlingsforløb med tilfælde af fast antipsykotisk monoterapi.

- Kataton adfærd
- Negative symptomer

Patient med paranoid skizofreni Uddrag af journalnotat:

Patienten indleder samtalen med at sige, at lyset ikke må tændes, fordi patienten da vil eksplodere. Kan ikke realitetskorrigeres i dette verbalt. Personalet foreslår patienten at forlade stuen, så lyset kan tændes hvorved patienten kan se at det ikke er farligt. Patienten accepterer dette og går ud på gangen, men er tydeligt skræmt da lyset tændes inde på stuen (patienten har ikke noget imod at stå på gangen, hvor lyset også er tændt). Patienten gemmer sig bag væggen og rækker en hånd ind og slukker for lyskontakten. Patienten kommer derefter ind på stuen igen...
Patienten er fortsat svært psykotisk med vrangforestillinger. „der er hul på hjernen", „jeg går i opløsning", „jeg er vred indeni". Patienten siger gentagne gange til personalet: „I kommer da til at ligne jer selv igen, ikke? jeg kan ikke kende jer". Patienten ser flere gange ned af sig selv og siger: „Hvordan er det jeg ser ud efterhånden"?
Patienten mener der er gift i kødet til middag og spytter det ud.

3.2.2 Skizotypisk sindslidelse (F21)

Den skizotypiske sindslidelse er nært slægtet med skizofreni i den forstand at patienten i sin adfærd fremtræder sær og med tendens til tilbagetrækning, ligesom der optræder forbigående psykotiske episoder i sygdomsforløbet. Et væsentligt kriterium for at få stillet diagnosen er dog, at patienten på intet tidspunkt har tilfredsstillet kriterierne for at få stillet diagnosen skizofreni [WHO, 1993].

Patient med skizotypisk sindslidelse Uddrag af journalnotat:

Patienten blev indlagt efter tiltagende isolationstendens gennem måneder. Patienten accepterede indlæggelse på grund af mange belastende og onde tanker, som patienten beskriver som påvirkning fra dæmoner. Patienten har særdeles vanskeligt ved kontakt til mænd og forlanger, at der ikke må være mænd blandt kontaktpersonerne. "Det giver kun negative følelser og tanker".

Patientens tanker beskrives som dæmoniske stemmer. Patienten har kontakt til de døde og siger, at det ikke er hallucinationer men åndeligt. De døde opfordrer patienten til at gøre noget voldeligt ved tilfældige forbipasserende på gaden, og til at tage sit eget liv. Patienten må bruge al sin energi på at holde disse tanker væk.

3.2.3 Kronisk paranoide psykoser (F22)

Paranoia kommer af det græske ord for forrykthed. Symptomerne på de kroniske paranoide psykoser er karakteriseret ved perioder af minimum 3 måneders varighed med forekomst af nonskizofrene vrangforestillinger. Til forskel for de skizofrene vrangforestillinger er paranoikerens overbevisninger kendetegnet ved et persekutorisk tema, der af omverdenen blot opleves som en absurditet og som åbenlyst ukorrekt. [Birkeland, 2007].

Patient med kronisk paranoid psykose Uddrag af journalnotat:

Patienten sidder på sengekanten. Er både udtalt vred og meget forpint. Fortæller at nogen er kommet ind i værelset og har stjålet nogle vigtige papirer fra det låste skab. Adspurgt om hvem har gjort det, svarer patienten. at det er en mand som ejer flere hoteller i København og som har sendt nogen afsted for at stjæle papirerne.

3.2.4 Skizoaffektiv tilstand (F25)

Sygdomsbilledet ved den skizoaffektive tilstand er kendetegnet ved symptomer af både affektiv og skizofren karakter. Diagnosen underinddeles i en manisk type og en depressiv type [WHO, 1993]. På dansk omtales tilstanden også som en blandingspsykose.

Patient med skizoaffektiv tilstand Uddrag af journalnotat:

Patienten har igennem en god uges tid igen været tiltagende irritabel, tidvis højtråbende. Patientens nattesøvn er igen kompromitteret: „Jeg bliver manisk om natten". Patienten har i nat skrevet et digt og to sange. I samtale nævner patienten at have været til interview i 2 timer, hvilket var anstrengende. Patienten klager over at få én diagnose det ene sted og siden en anden diagnose på afdelingen.
Patienten fremtræder med et vist forsænket stemningsleje, svarer lidt stereotypt og automatisk men uden latenstid. Patienten virker ikke umiddelbart produktivt psykotisk under samtalen, men refererer til tidligere at have hørt stemmer.

3.3 Behandling med antipsykotika

En væsentlig udfordring ved behandlingen af de skizofrene lidelser er at ætiologien[2] endnu ikke er fuldt klarlagt. I tilfælde med skizofreni er der indikationer på at den væsentligste sygdomsprogression sker indenfor 5 år af den første psykotiske episode, hvilket [Sanger *et al.*, 1999] fremhæver som et muligt tegn på en underliggende patofysiologisk proces. Dette tydeliggører naturligvis nødvendigheden af hurtig og virksom intervention. Og forsøg viser da også at sådanne tiltag tidligt i sygdomsforløbet i tilfælde med skizofreni kan forbedre patientens prognose [Frangou and Byrne, 2000]. Selvom de kliniske interventionsmetoder har udviklet sig drastisk fra fremkomsten af det første antipsykotikum i 1952[3] er der fortsat to væsentlige problematikker ved den antipsykotiske behandling, som man i dag er i stand til at tilbyde: begrænset effekt samt bivirkninger.

Begrænset effekt

Behandling med antipsykotika er ikke kurativ[4] og særligt de skizofrene lidelser giver sig udslag i en livslang invaliderende tilstand for de berørte patienter. I tilfælde med

[2] Ætiologi beskriver årsagen til en given lidelse med identifikation af dens disponerende faktorer.

[3] Se afsnit 3.3.2.

[4] Det at en behandling er kurativ betyder at patienten efter behandlingens afslutning vil være helbredt.

behandling med typiske antipsykotika[5] er der i følge [Jibson and Tandon, 1998] tale om, at 30% af de psykotiske patienter ikke vil opleve en signifikant terapeutisk effekt af behandlingen. Fremkomsten af atypiske antipsykotika har dog forbedret behandlingen af særligt de skizofrene lidelser væsentligt selvom den forbedrede effekt, der tilskrives disse antipsykotika er mål for en del debat, særligt med hensyn til de medfølgende birvikninger [Gaebel *et al.*, 2007].

Bivirkninger

Behandling med antipsykotika har en række uønskede effekter også kendt som bivirkninger. Da virkningsmekanismen[6] for antipsykotika har betydning for både den ønskede effekt ligsåvel som de uønskede effekter og effekten er dosisafhængig, vil det ofte være fremkomsten af bivirkninger, der er den begrænsende faktor i forhold til udfaldet af behandlingen [Sekine *et al.*, 1999].
Behandlingseffekten ved den antipsykotiske behandling vil være et gennemgående tema for de følgende afsnit. Bivirkningerne ved den antipsykotiske behandling vil blive yderligere gennemgået i afsnit 3.3.3.

3.3.1 Antipsykotika

Navngivning

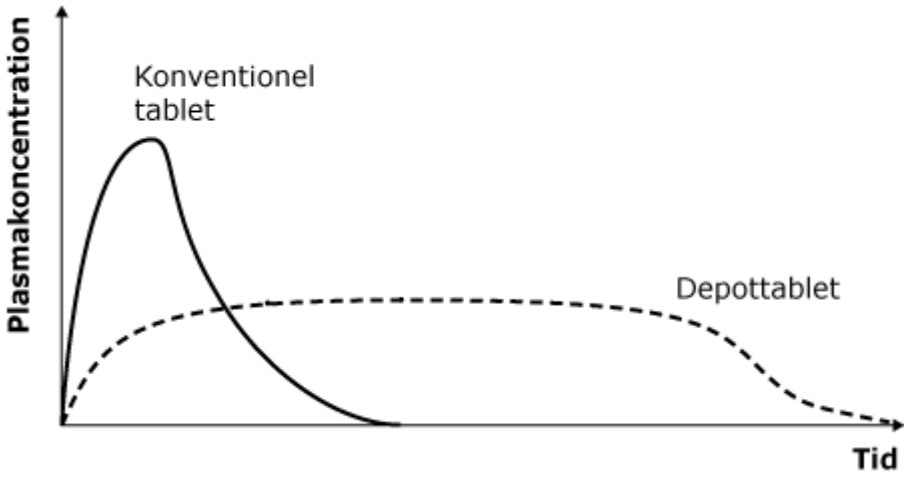
Et antipsykotikum er et præparat, der anvendes til behandling af de psykotiske symptomer, hvis effekt skyldes aktivstoffets kemiske struktur [Sekine *et al.*, 1999]. Præparatetnavnet er det *handelsnavn*, som den enkelte medicinalvirksomhed markedsfører stoffet og dets form under (tablet, væske til injektion osv.). Udover handelsnavnet vil præparatet også være tildelt et *generisk navn* på baggrund af aktivstoffets kemiske struktur. Man støder ofte på flere konkurrende præparater (med forskellige handelsnavne), der indeholder samme aktivstof og derfor deler samme generiske navn.
I specialet anvendes både handelsnavnet og præparaternes generiske navne. Den anvendte navngivning vil dog fremgå af teksten ligesom handelsnavnet altid vil være anvendt med stort begyndelsesbogstav.

Administration

Antipsykotiske præparater forefindes både på konventionel doseringsform og i depotform. De konventionelle doseringsformer inkluderer tabletter, kapsler eller dråber til oral administration, hvor patienten synker præparatet efter indtagelse gennem munden. Optagelse af præparatets indholdsstof sker efterfølgende ved *absorption* i mave-tarmkanalen. Absorption henviser til præparatets optagelse i kroppens systemiske kredsløb, det vil sige blodbanen. Via blodet overføres præparatets indholdsstof til hjernen [Alavijeh *et al.*, 2005].

[5] Se afsnit 3.3.2.
[6] Se afsnit 3.3.1.

Depotformen består primært af væske til intramuskulær injektion. Et enkelt depotpræparat (Semap) findes dog på tabletform. Ved intramuskulær injektion indsprøjtes patienten med præparatet på væskeform i en fedtholdig opløsning, der medvirker til præparatets binding til muskelvæv [Patel and David, 2005]. En sådan administration medfører, at præparatet absorberes over længere tid, hvorved en nogenlunde konstant plasmakoncentration opretholdes [Patel and David, 2005].



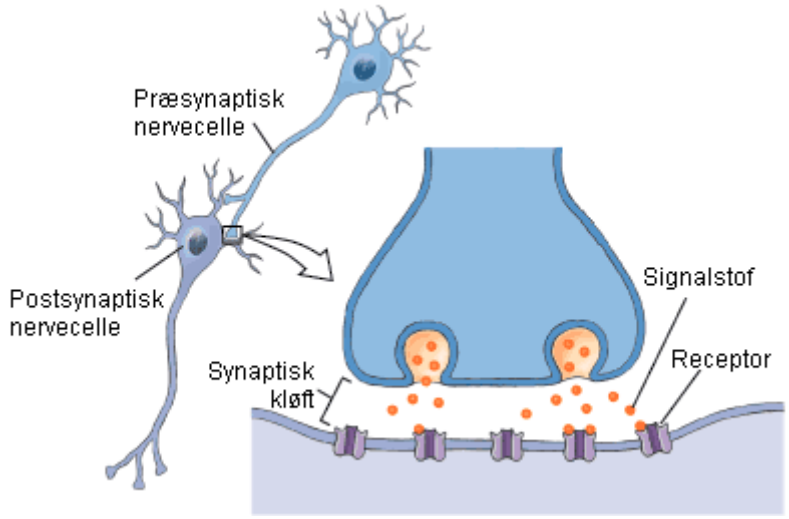
Figur 3.2: Plasmakoncentrationer ved heholdsvis konventionel dosering og depotadministration. Frit efter [Pedersen *et al.*, 2008]

På figur 3.2 er skitseret plasmakoncentrationen ved henholdsvis den konventionelle doseringsform og administration af et depotpræparat. Figuren illustrerer, hvordan plasmakoncentrationen af den konventionelle tablet gradvist øges, til den når sit maksimum, imens depottabletten opretholder en længerevarende fast plasmakoncentration. For begge administrationsformer gælder, at koncentrationen falder, når præparatets nedbrydning begynder at overstige dets absorption. Plasmakoncentrationen varierer naturligvis mellem de forskellige præparater og til forskellige doser, men overordnet set kan varigheden af den konventionelle doseringsform udtrykkes i timer eller i dage, imens varigheden af depotformen kan udtrykkes i uger [Patel and David, 2005].

Virkningsmekanisme

Efter absorption af præparatet er det aktivstoffets kemiske struktur, der sætter det i stand til at binde sig til strukturer i patientens neuroner (nerveceller) kaldet receptorer. Den enkelte receptor fungerer som modtagerenhed ved neurotransmissioner (signaloverførsler mellem hjernens nerveceller). Rummet mellem to nerveceller betegnes den *synaptiske kløft* og ved neurotransmission benyttes betegnelsen den *præsynaptiske nervecelle* til at angive det neuron som signalet udgår fra mens modtagerneuronet betegnes den *postsynaptiske nervecelle*. Figur 3.3 illustrerer neurotransmissionen over den synaptiske kløft, hvor signalstoffer frigøres fra den præsynaptiske nervecelle for

derefter at binde sig til receptorer i den postsynaptiske nervecelle.



Figur 3.3: Neurotransmission med frigørelse af signalstoffer fra den præsynaptiske nervecelle og optagelse af disse stoffer ved hjælp af receptorer i overfladen af den postsynaptiske nervecelle. Figur af [Morris and Maisto, 2000].

Ved at binde sig til neuronernes receptorer blokerer antipsykotika for hjernens signalstoffer hvorved neutransmissionen hæmmes, hvilket er grundlaget for den antipsykotiske effekt. Hjernens neuroner indeholder en række forskellige receptorer, og selvom flere af disse er mål for især nyere (atypiske) antipsykotika, er det særligt den postsynaptiske D2-receptor, som antipsykotika har *affinitet* for [Lambert and Castle, 2003]. Affinitet benyttes som mål for tilbøjeligheden af stoffet til at binde sig til den pågældende receptor. At affiniteten er høj medfører, at stofferne i høj grad binder sig til D2-receptorerne og dermed blokerer for optagelsen af signalstoffet dopamin. En sådan blokade betegnes som en antagonistisk effekt, hvorfor antipsykotika siges at fungere som *antagonister* for D2-receptorer.
Udviklingen af antipsykotika har medført, at en lang række præparater med forskellige *receptorprofiler* er dukket op på markedet [Lieberman *et al.*, 2005]. Receptorprofilen beskriver de receptorer som det enkelte antipsykotikum har affinitet for. Jeg har i dette afsnit valgt at tage udgangspunkt i D2-receptoren, idet den som nævnt er mål for samtlige af de antipsykotika, der findes på markedet i dag. Selvom især atypiske antipsykotika har højere affinitet for andre receptorer [Lieberman *et al.*, 2005], tjener det valgte eksempel til at beskrive den generelle virkningsmekanisme ved antipsykotika.

3.3.2 Klassifikation af antipsykotika

Antipsykotika opdeles overordnet set i typiske og atypiske antipsykotika [Lewis and Lieberman, 2000]. Opdelingen sker primært på baggrund af præparaternes virkning, det vil sige deres receptorprofiler. Typiske antipsykotika betegnes også førstegenerations-

antipsykotika og de atypiske som andengenerations-antipsykotika [Lieberman *et al.*, 2005]. Denne navngivning hænger dog til dels sammen med præparaternes udvikling. Det første antipsykotikum, Chlorpromazin, så dagens lys i 1952 og årene efter udvikledes flere antipsykotika dog alle med høj affinitet for D2-receptoren [Lambert and Castle, 2003]. Clozapin, der kom på markedet i USA i 1990 [Cobaugh *et al.*, 2007] markerede overgangen til de atypiske antipsykotika med langt bredere receptorprofiler end de typiske antipsykotika [Lieberman *et al.*, 2005].
Kendskabet til receptorprofilerne for antipsykotika stammer blandt andet fra såkaldte *in vitro*[7] forsøg. Og skønt udtrykket af receptorerne er individafhængige på samme måde som andre faktorer af betydning for farmakodynamikken[8], danner kendskabet til receptorprofilerne baggrund for forståelsen af antipsykotikas både ønskede og uønskede effekter, selvom den grundlæggende bagvedliggende mekanisme fortsat er ukendt [Dazzan *et al.*, 2005].

Typiske antipsykotika

Disse antipsykotika gik før introduktionen af atypiske antipsykotika under betegnelsen neuroleptika, og er opstillet i tabel 3.4 med angivelse af både præparatnavn og generisk navn.
Tre artikler om kliniske forsøg med præparaterne haloperidol (Serenase) [Okasha and Tewfik, 1964] og [McCreadie and MacDonald, 1977] og chlorprothixen (Truxal) [Remvig and Sonne, 1961] fra perioden før indtroduktionen af de atypiske antipsykotika er fremstillet i tabel 3.5. Clorpromazin blev som det første neuroleptika udgangspunktet for de beskrevne kliniske undersøgelser.
Generelt afslører forsøgene ingen entydig forbedring for den samlede patientpopulation. Dette illustrerer at præparatets virkning (også omtalt som patients *respons* på præparatet) er meget individafhængig. Selvom et mindre udsnit af forsøgenes patientkohorte oplever signifikant forbedring af tilstanden er det værd at bemærke, at de personer der udviser et fravær af en forbedret klinisk effekt og/eller væsentlig højere eller alvorligere bivirkningsforekomst end ved patientens tidligere behandling, allerede er udtrådt af patientkohorten ved forsøgets afslutning.
Den inhibitoriske effekt af de typiske antipsykotika på dopamreceptorer i hjernen udgør grundlaget for den antipsykotiske effekt som tidligere beskrevet. Idet dopaminerge receptorer imidlertid også er tilstedeværende i det ekstrapyramidale system, der er styrende for bevægeapparatet har den høje affinitet for D2-receptorer for de typiske antipsykotika en række velbeskrevede bivirkninger kendt som ekstrapyramidale symtomer (EPS). Sådanne symptomer kommer til udtryk som stivhed, rigiditet og kramper i musklerne. Observationer som også er gjort af [Okasha and Tewfik, 1964] og [McCreadie and MacDonald, 1977] og i væsentlig mindre grad af [Remvig and Sonne, 1961]. At [Remvig and Sonne, 1961] kun observerer få EPS skyldes givet vis, at Truxal er højdosis-præparat, der opnår en højere antipsykotisk effekt ved lavere dosis

[7] In vitro er latin for *i glasset* og in vitro forsøg er kort skitseret forsøg med udvalgte celler eller vævstyper (eksempelvis hjernevæv), der udtrykker en eller flere udvalgte receptorer.
[8] Farmakodynamikken beskriver lægemidlets virkning på organismen.

end tilfældet er for lavdosis-præparatet Serenase. Både [Remvig and Sonne, 1961] og [McCreadie and MacDonald, 1977] oplever desuden væsentlige tegn på leverskade ved behandlingen.
De generelle bivirkninger ved antipsykotika er gennemgået i afsnit 3.3.3.

Handelsnavn	**Generisk navn**
Lavdosis	
Fluanxol	flupentixol
Siqualone	fluphenazin
Serenase	haloperidol
Orap	pimozid
Middeldosis	
Neulactil	periciazin
Trilafon	perphenazin
Stemetil	prochlorperazinmaleat
Cisordinol	zuclopenthixolacetat
Højdosis	
Truxal	chlorprothixen
Nozinan	levomepromazin
Buronil	melperon
Dipiperon	pipamperon
Dogmatil	sulpirid

Tabel 3.4: Typiske antipsykotika med angivelse af både præparatets handelsnavn samt dets generiske navn. Opdelingen i lav-, middel- og højdosis præparater henviser til den nødvendige dosisstørrelse for opnåelse af den antipsykotisk effekt. Lavdosis antipsykotika kræver således en lavere dosis end højdosis antipsykotika for opnåelse af den antipsykotiske effekt [Pedersen *et al.*, 2008].

Atypiske antipsykotika

Præparatnavne for de atypiske antipsykotika er oplistet i tabel 3.7 med angivelse af både handelsnavne og generiske navne. En række kliniske undersøgelser af disse præparater er angivet i tabel 3.6 og benyttes som udgangspunkt for gennemgangen af denne gruppe antipsykotika.
De atypiske antipsykotika har ved sammenligning med de typiske antipsykotika en signifikant bedre virkning [Sanger *et al.*, 1999] med fremhævelse af effekten på de negative symptomer [Rubio *et al.*, 2006a]. Ikke overraskende medfører den lavere affinitet for D2-receptorer et mindre udtryk af EPS [Sanger *et al.*, 1999], [Möller *et al.*, 2008] og [Schooler *et al.*, 2005] og generelt færre bivirkninger [Peuskens, 1995]. Ved sammenligning mellem forskellige atypiske antipsykotika er forskellen dog minimal, hvilket som ved tilfældet med de typiske antipsykotika, kan ses som et tegn på de individuelle forskelle i sygdomsudtrykket hos patienterne.

Tabel 3.5: Kliniske forsøg med typiske antipsykotika. Rf (resultatfortolkning) angiver om der ved undersøgelse af det enkelte præparat versus et andet, er tale om en signifikant forbedring givet ved artiklens konklusion(er). +1 angiver signifikant forbedring mens 0 angiver fravær af signifikant forbedring.

Artikel	Undersøgelse	Rf.	Resultater
[Remvig and Sonne, 1961]	**Chlorprothixen versus chlorpromazin:** 163 blandede psykiatriske patienter randomiseres til chlorprothixen eller chlorpromazin i et dobbelt-blindet forsøg med gradvis dosisøgning til der observeres en effekt.	0/+1	Psykotiske patienter opnåede enten samme eller bedre effekt med Chlorprothixene med færre indrapporterede bivirkninger i løbet af 7 uger.
[Okasha and Tewfik, 1964]	**Haloperidol versus placebo:** 69 kronisk syge psykotiske patienter som ikke responderer på anden behandling randomiseres i et dobbelt-blindet forsøg til enten placebo eller haloperidol.	0/+1	Sammenlignet med tidligere behandling er haloperidol ikke bedre men er dog signifikant bedre end placebo. Tre patienter havde særligt gavn af behandlingen hvoraf den ene havde særdeles godt respons på positive symptomer og blev udskrevet.
[McCreadie and MacDonald, 1977]	**Haloperidol versus chlorpromazin:** 20 kronisk skizofrene mandlige patienter som ikke responderer på tidligere behandling randomiseres i et enkelt-blindet forsøg til enten haloperidol eller chlorpromazin.	+1	Ved høj dosis haloperidol viste denne gruppe signifikant bedring i forhold til chlorpromazin-gruppen uden tilstedeværelse af flere bivirkninger.

Tabel 3.6: Kliniske forsøg med atypiske antipsykotika. Rf (resultatfortolkning) angiver om der ved undersøgelse af det enkelte præparat versus et andet, er tale om en signifikant forbedring givet ved artiklens konklusion(er). +1 angiver signifikant forbedring, 0 angiver fravær af signifikant forbedring mens -1 angiver signifikant ringere effekt. [1]) EPS er forkortelsen for ekstrapyramidale symptomer, som kommer til udtryk som rigiditet, stivhed og krampe i musklerne.

Artikel	Undersøgelse	Rf.	Resultater
[Sanger *et al.*, 1999]	**Olanzapin versus haloperidol:** 83 første-episode psykotiske patienter i behandling med enten olanzapin eller haloperidol.	+1	Patienter behandlet med olanzapin opnåede en signifikant reduktion i psykotiske symptomer. Haloperidol havde tilgengæld signifikant flere EPS[1].
[Nelson *et al.*, 2001]	**Olanzapin:** 7 patienter med depressiv psykose deltog i et pilotstudie for at vurdere monoterapi med olanzapin i 10 uger.	+1	Der var en signifikant forbedring af både depressive og psykotiske symptomer ved sammenligning af vurderinger fra starten og til undersøgelsens afslutning.
[Riedel *et al.*, 2007]	**Olanzapin versus quetiapin:** 52 skizofrene patienter randomiseres i et dobbel-blindet forsøg til enten olanzapin eller quetiapin i 8 uger.	-1/0	Begge grupper opnåede kognitiv forbedring. Quetiapin-gruppen opnåede dog en signifikant bedre score ved opmærksomhedsvurdering.
[Curtis *et al.*, 2008]	**Risperidon:** 842 psykotiske patienter med høj forekomst af negative symptomer selekteres fra patientkohorten i StoRMi-forsøget hvor alle modtager risperidon.	+1	PANS-skala for positive og negative symptomer benyttes for den samlede StoRMi-kohorte. Subgruppen på 842 opnår signifikant større reduktion af negative symptomer.
[Möller *et al.*, 2008]	**Risperidon versus haloperidol:** 289 første-episode patienter randomiseres i et dobbel-blindet forsøg til enten risperidon eller haloperidol i 8 uger.	0/+1	Der var ingen signifikant forskel på den antipsykotiske effekt i de to grupper. Haloperidol-gruppen havde dog dobbelt så høj chance for at få EPS[1].
[Rubio *et al.*, 2006b]	**Risperidon versus zuclopenthixol:** 66 skizofrene patienter med forskelligt misbrug deltog i et åbent kryds-over forsøg med henholdsvis risperidon og zuclopenthixol.	+1	Risperidon-gruppen havde på begge sider af krydset et lavere misbrug ligesom kompliansen var højere. Generelt fremkom risperidon-gruppen med færre negative symptomer.

Fortsættes på næste side...

Tabel 3.6 – Fortsat

Artikel	Undersøgelse	Rf.	Resultater
[Peuskens, 1995]	**Risperidon versus haloperidol:** 1362 kronisk skizofrene patienter fra centre i 15 lande randomiseres til en af fem doser risperidon eller haloperidol.	0/+1	Signifikant flere bivirkninger i haloperidolgruppen dog ingen signifikant højere forekomst af ønskede virkninger mellem grupperne.
[Rubio *et al.*, 2006a]	**Risperidon versus zuclopenthixol:** 115 skizofrene patienter med forskelligt misbrug deltog i et åbent forsøg hvor patienter enten modtog risperidon eller zuclopenthixol.	+1	For risperidon-gruppen observeredes et lavere misbrug samt en reduktion af negative symptomer.
[Schooler *et al.*, 2005]	**Risperidon versus haloperidol:** 555 første-episode patienter randomiseres i et dobbel-blindet forsøg til enten risperidon eller haloperidol.	0/+1	Der optrådte signifikant flere EPS[1]symptomer i haloperidol-gruppen mens prolactin-øgning var størst i risperidon-gruppen. Der var ingen signifikant bedre behandlingseffekt mellem grupperne dog med størst tilbagefald i haloperidol-gruppen.

Handelsnavn	**Generisk navn**
Solian	amisulprid
Abilify	aripiprazol
Clozapin	clozapin
Leponex	clozapin
Zyprexa	olanzapin
Invega	paliperidon
Seroquel	quetiapin
Risperdal	risperidon
Serdolect	sertindol
Zeldox	ziprasidon
Semap	penfluridol

Tabel 3.7: Atypiske antipsykotika med angivelse af både præparatets handelsnavn samt dets generiske navn [Pedersen *et al.*, 2008].

3.3.3 Bivirkninger

Ligesom den antipsykotiske effekt kan tilskrives det enkelte præparats binding til receptorer i neuronets overflade, medfører de samme receptorbindinger også en række uønskede effekter. Det enkelte præparats receptorprofil er derfor også afgørende for de fremkomne bivirkninger, hvilket har ført til den parallele betegnelse præparaternes bivirkningsprofiler. Kendskab til præparaternes receptor- og bivirkningsprofiler kan dermed støtte klinikeren i at målrette den antipsykotiske behandling, for at opnå så gavnlig en virkning af behandlingen som muligt samtidig med, at forekomsten af bivirkninger forsøges minimeret.

Bivirkninger på centralnervesystemet

Bivirkninger på centralnervesystemet ses hyppigst ved behandling med typiske antipsykotika og kommer blandt andet til udtryk som *dyskinesi*, *ekstrapyramidale symptomer* (EPS) og *sedation* [Arana, 2000].

Dyskinesier. Dyskinesier betegner ufrivillige abnorme bevægelser, der ofte ses omkring munden. Dyskinesier opdeles i *akutte* og *tardive*, der er den kliniske term for henholdsvis tidlig og sen optræden af symptomerne. Dyskinesier forekommer hyppigt som tardive og ses sjældent akut. Ligesom ved parkinsonisme kan tilstanden være til stede efter behandlingsophør. Langvarig behandling giver desuden øget risiko for, at tilstanden skal blive permanent [Arana, 2000], hvilket i klinisk terminologi betegnes som *irreversibel*.

Ekstrapyramidale symptomer. Ekstrapyramidale symptomer (EPS) inkluderer *akatisi*, det vil følelsen af rastløshed, der kommer til udtryk som uro og manglende evne til at sidde stille; *akinesi*, det vil sige nedsat eller forsinket bevægelighed samt *dystoni*, som er karakteriseret ved længerevarende ufrivillige trækninger i enkelte eller flere muskelgrupper som typisk optræder tidligt i behandlingen eller ved dosisøgning. Dystoni kan dog bringes til nærmest øjeblikkelig ophør ved hjælp af bivirkningsmedicin [Arana, 2000].

Sedation. Typiske antipsykotika har en *sederende* virkning der kommer til udtryk som træthed eller døsighed. Denne effekt er årsag til brug af typiske antipsykotika til sedation af udfarende eller aggressive patienter. Den sederende effekt er størst ved højdosis-præparaterne[9] arana$_2$000.

Endokrine forstyrrelser

Endokrine eller hormonelle forstyrrelser der skyldes et forhøjet prolaktinniveau ses ved både typiske og atypiske antipsykotika [Arana, 2000] og [Serretti *et al.*, 2004]. Prolaktinproduktionen spiller en rolle hos både mænd og kvinder idet den hjælper med

[9] Se tabel 3.4.

at regulere dannelsen af kønshormoner. Signalstoffet dopamin indgår foruden i neurotransmissionen også i reguleringen af prolaktin. Ved visse antipsykotika-anvendelser begrænses virkningen af dopamin imidlertid, hvilket kan føre til en ureguleret produktion af prolaktin, der hæmmer dannelsen af kønshormoner. Symptomerne hos kvinder er udebleven menstruation, mælkeproduktion og galaktoré (mælkeudskillelse). For både mænd og kvinder fremstår symptomer med nedsat libido (sexualdrift) og osteoporose (knogleskørhed) [Serretti *et al.*, 2004].

Metaboliske forstyrrelser

Metabolismen (eller stofskiftet) benyttes som samlebetegnelse for kroppens energiforbrugende processer. De metaboliske forstyrrelser fremkommer derfor som patofysiologiske defekter udtrykt ved vægtøgning, diabetes mellitus, dyslipidæmi og det metaboliske syndrom [Leitão-Azevedo *et al.*, 2006]. Metaboliske forstyrrelser forekommer både ved behandling med typiske som atypiske antipsykotika.

Vægtøgning. Vægtøgning er for de atypiske antipsykotikas vedkommende især associeret med præparaterne clozapin (Clozapin og Leponex) og olanzapin (Zyprexa) [Serretti *et al.*, 2004].

Diabetes mellitus. Diabetes mellitus betegnes også type 2 sukkersyge og skyldes insulinresistens. Insulinresistens referer til hormonet insulins manglende evne til at transportere sukker ind i kroppens celler. Insulinresistens og glukoseintolerans optræder både som forstadium til og under sygdomsforløbet ved diabetes mellitus. Glukoseintolerans er en tilstand, hvor kroppens celler i meget lille grad er i stand til at optage sukker. Både insulinresistens og glukoseintolerans medfører hyperglykæmi (forhøjet blodsukker). Symptomerne er blandt andet osmotisk diurese (øget vandladningstendens), hypovolæmi (væsketab), hypovolæmisk hypotension (nedsat blodtryk på grund af væskemangel), kvalme, opkastninger, svimmelhed og chok. Igen er clozapin og olanzapin væsentligste bidragerydere til denne bivirkning sammen med quetiapin (Seroquel) [Serretti *et al.*, 2004].

Dyslipidæmi. Lipidstofskiftet benyttes som betegnelse for de processer der blandt andet regulerer omsætning af fedtstoffer (lipider) og udtrykket af fedtstoffer i blodkarrene. Defekter i lipidstofskiftet også kaldet dyslipidæmi viser sig blandt andet ved forhøjet kolesteroltal, der disponerer patienter for arteriosklerose (åreforkalkning) og andre kardiovaskulære lidelser som hypertension (forhøjet blodtryk) [Kannabiran and Singh, 2008].

Det metaboliske syndrom. Det metaboliske syndrom betegner tilstedeværelsen af flere samtidige risikofaktorer for kardiovaskulære lidelser i form af insulinresistens, dyslipidæmi, hypertension eller glukoseintolerans [Kannabiran and Singh, 2008].

Antikolinerge bivirkninger

Antikolinerge bivirkninger skyldes blokade af neuronernes kolinerge og noradrenerge receptorer. Begge receptorer er af væsentlig betydning for funktionen af det parasympatiske nervesystem, der regulerer kroppens *i-hvile*-funktioner. Antipsykotikas blokering af disse receptorer betegnes også den antikolinerge effekt. Symptomerne herpå er mundtørhed, uskarpt syn, forstørrede pupiller, vandladningsbesvær og obstipation (forstoppelse) [Jibson and Tandon, 1998].

3.4 Antipsykotiske behandlingsforløb

Specialets definition af behandlingsforløb, som perioden med samme kontinuerlige medicinske behandling[10], muliggør en tydeliggørelse af de farmakologiske interventioner, som er helt centrale for psykosebehandlingerne.
Påbegyndelsen af interventionen er karakteriseret ved indtræden i behandlingsforløbet og ophøret betegnet som forløbets behandlingsudfald. Idet der i specialet ses bort fra både administrationsmetode og dosisstørrelse skelnes alene mellem monoterapi og polyfarmaci for det pågældende behandlingsforløb.
Figur 3.4 der illustrerer tilstedeværelsen af to på hinanden følgende behandlingsforløb, vil blive anvendt i afsnittets beskrivelse af antipsykotiske behandlingsforløb. På figuren er behandlingsforløbene afbrudt af en periode med polyfarmaci på grund af behandlingsskifte fra præparat A til præparat B.

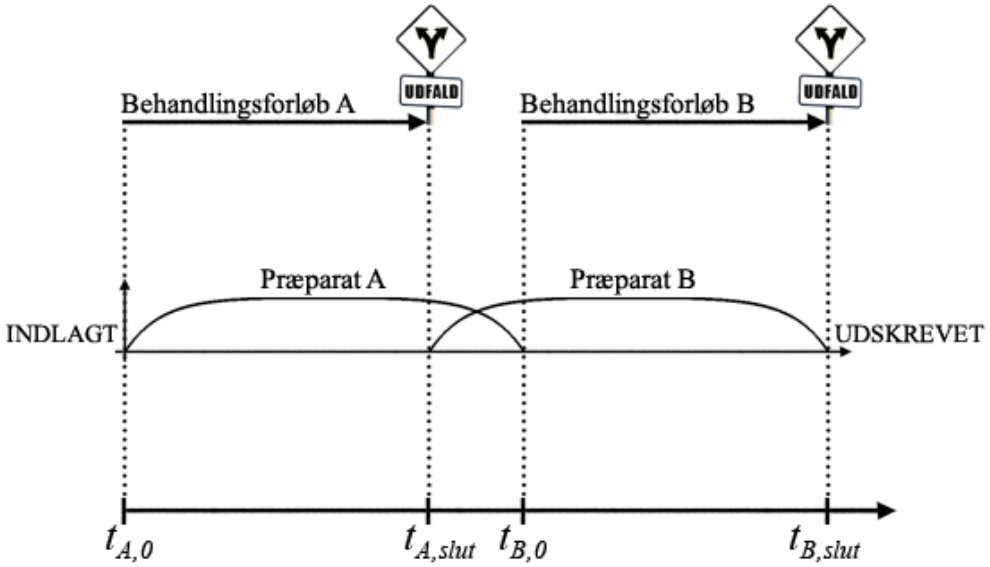
3.4.1 Præparatvalg

Præparatvalg kendetegner påbegyndelsen af interventionen karakteriseret ved tilføjelsen af en ny ordination. På figur 3.4 er der således tilfælde af præparatvalg til tidspunkterne $t_{A,0}$ og $t_{A,slut}$.
Den generelle konsensus indenfor psykiatrien både i og uden for Europa for førstevalg af præparater til behandlingen af skizofreni lyder på valg af et atypisk antipsykotikum [Mauri *et al.*, 2005], [Apiquian *et al.*, 2004] & [Trifiro *et al.*, 2005]. Valg af et atypisk antipsykotikum har den fordel, at der kun i begrænset omfang forekommer ekstrapyramidale symptomer (EPS) ved behandlingen [Sanger *et al.*, 1999], ligesom flere kliniske undersøgelser peger på en bedre behandlingseffekt ved behandling med atypiske antipsykotika i forhold til behandling med typiske antipsykotika [Riedel *et al.*, 2007], [Curtis *et al.*, 2008] & [Rubio *et al.*, 2006a].

3.4.2 Monoterapi

På figur 3.4 er der tilfælde af monoterapi med præparaterne A og B henholdsvis i perioderne $t_{A,0}$ til $t_{A,slut}$ og $t_{B,0}$ til $t_{B,slut}$.
Monoterapi anbefales som hovedregel til enhver antipsykotisk behandling [Faries *et al.*, 2005]. Den væsentligste årsag hertil er behandlingsevidensen. Erfaringer høstet

[10] Se afsnit 2.4.5.



Figur 3.4: To på hinanden følgende behandlingsforløb baseret på ordination af præparat A ved tidspunktet $t_{A,0}$ og præparat B ved tidspunktet $t_{A,slut}$. Behandlingsforløb A (med præparat A i monoterapi) afsluttes med behandlingsudfaldet behandlingsskifte, som adskiller behandlingsforløbenes monoterapi med tilfælde af polyfarmaci i perioden $t_{A,slut}$ til $t_{B,0}$. Indtræden i behandlingsforløb B (med præparat B i monoterapi) sker ved tidspunktet $t_{B,0}$ og ophører ved behandlingsudfaldet udskrivning ved tidspunktet $t_{B,slut}$.

fra kliniske forsøg siden fremkomsten af chlorpromazin i 1952 er væsentligste bidragyder til forståelsen for behandlingseffekten ved antipsykotika og udgør grundlaget for blandt andet vejledningen til behandling med antipsykotika i Region Hovedstaden [Glenthøj *et al.*, 2008], som Psykiatrisk Center Sct. Hans hører under.
Der er desuden evidens for en højere grad af *komplians* ved monoterapi end ved polyfarmaci [Morrato *et al.*, 2007] & [Love, 2002]. Komplians repræsenterer patientens velvillighed til at indtage sin medicin. Andre faktorer som bivirkningsforekomster har dog også betydning for kompliansen. Således vil præparater med færre bivirkninger ofte medføre højere komplians end andre præparater [Love, 2002] ligesom gentagende indlæggelser og sygdomstilbagefald medfører en lavere grad af komplians [Love, 2002].

3.4.3 Polyfarmaci

Generelt skelnes mellem *kortvarig* og *langvarig* polyfarmaci [Kreyenbuhl *et al.*, 2007]. Kortvarig polyfarmaci er kendetegnet ved skifte mellem to præparater, imens langvarig polyfarmaci skyldes fast behandling med flere samtidige præparater. På figur 3.4 er behandlingsforløbene således adskilt af kortvarig polyfarmaci.
I specialet skelnes ikke mellem den kortvarige og langvarige form. Selvom kritikken der rettes mod brug af polyfarmaci vedrører den langvarige form, betragtes udfaldet behandlingsskifte fortsat som et negativt udfald af behandlingsforløbet. Det skal dog bemærkes, at kortvarig polyfarmaci er generelt accepteret [Miller and Craig, 2002]

& [Lader, 1999] idet den antipsykotiske effekt af at givent præparat normalt først indtræder flere uger efter påbegyndelse af ordinationen, hvorfor psykosebehandlingen således bedst opretholdes ved ordination af et nyt præparat nogle uger før ophør af det foregående præparat [Masand, 2005].
Den kritik der samler sig om polyfarmaci er primært begrundet i følgende to forhold:

1. Manglende evidens
2. Risiko for ukendte lægemiddelinteraktioner

Desuden nævnes svigtende komplians, højere økonomiske omkostninger og øget risiko for

Manglende evidens

Evidensen for anvendelsen af polyfaramaci er stort set ikke eksisterende hvorfor de kliniske retningslinier, der nævner denne mulighed, beskriver den som en *sidste* mulighed efter gentagne fejlslagne monoterapiforløb [Faries *et al.*, 2005] der således kun anbefales *behandlingsrefraktære patienter*[11] [Love, 2002]. I tabel 3.8 er gengivet to kliniske forøg med stærkt behandlingsrefraktære patienter, der begge undersøger polyfarmaci med clozapin og risperidon. Selvom der i undersøgelsen af [Josiassen *et al.*, 2005] observeredes en signifikant forbedring af de negative symptomer er det dog langt fra nok til at skabe evidens for behandlingen i polyfarmaci.

Risiko for ukendte lægemiddelinteraktioner.

Idet samtidig anvendelse af flere antipsykotika øger risikoen for ukendte lægemiddelinteraktioner [Miller and Craig, 2002], skabes ikke alene tvivl om behandlingens udfald men i lige så høj grad bekymring om alvorlige og uforudsete bivirkninger ved behandlingen.

3.4.4 Behandlingsrationale for valg af polyfarmaci

[Freudenreich and Goff, 2002] fremhæver følgende *teoretiske* bevæggrunde for valg af polyfarmaci:

1. Øget blokade af D2-receptorer
2. Flere samtidige receptorbindinger der medfører:
   (a) forbedret antipsykotisk effekt
   (b) færre bivirkninger som følge af lavere nødvendig dosis for opnåelse af den antipsykotiske effekt.

[11] Behandlingsrefraktære patienter betegner de patienter der ikke har oplevet nogen effekt af flere forskellige antipsykotika.

Tabel 3.8: Kliniske undersøgelser af polyfarmaci med clozapin for behandlingsrefraktære psykotiske patienter. Rf (resultatfortolkning) angiver om der ved undersøgelse af præparatsammensætningen versus en anden sammensætning, er tale om en signifikant forbedring givet ved artiklens konklusion(er). +1 angiver signifikant forbedring mens 0 angiver fravær af signifikant forbedring.

Artikel	Undersøgelse	Rf.	Resultater
[Josiassen *et al.*, 2005]	**Clozapin + risperidon versus clozapin + placebo:** 40 behandlingsrefraktære skizofrene patienter i clozapinbehandling randomiseres i et dobbelblindet forsøg til placebo eller risperidon sammen med clozapin i 12 uger.	+1	Patienter i polyfarmaci-gruppen opnåede en signifikant reduktion af negative symptomer. 2 patienter i polyfarmaci-gruppen fik bivirkninger i form af mild akatisi. Der var ingen forskel på øvrige bivirkninger.
[Honer *et al.*, 2006]	**Clozapin + placebo versus clozapin + risperidon:** 68 behandlingsrefraktære skizofrene patienter i clozapinbehandling randomiseres i et dobbel-blindet forsøg til placebo eller risperidon sammen med clozapin.	0	Der blev ikke observeret en forbedring ved introduktion af polyfarmaci. Der var tegn på forhøjet risiko for glukose-regulationsdefekt ved polyfarmaci.

Blokade af D2-receptorer fremhæves generelt som et væsentligt udgangspunkt for den antipsykotiske effekt, da alle antipsykotika har enten større eller mindre affinitet for denne receptor. Øget blokade af denne receptor som følge af tilstedeværelsen af flere samtidige antipsykotika har derfor en hypotetisk bedre antipsykotisk virkning [Freudenreich and Goff, 2002].

Flere samtidige receptorbindinger som følge af polyfarmaci har hypotetisk set en additiv effekt i form af de enkelte præparaters dokumenterede effekter. Samtidig kan dosisstørrelsen for præparaterne holdes lavere end hvis de var givet i monoterapi hvorved bivirkningsudtrykket minimeres [Freudenreich and Goff, 2002].

Problemet er imidlertid, at ovennævnte bevæggrunde alene baserer sig på et teoretisk grundlag, hvilket også [Freudenreich and Goff, 2002] bemærker. Med dette in mente vil behandlingsforløb der udgøres af tilfælde af polyfarmaci således ikke være velbegrundet og rationalet for den valgte behandling således være sløret af den manglende evidens, der knytter sig til både behandlingen og dens udfald. Dette er årsagen til, at polyfarmaci ikke medtages i specialets undersøgelser. At brugen af polyfarmaci desuden kun beskrives som en *mulighed* for behandlingsrefraktære patienter [Love, 2002], gør at inddragelse af sådanne behandlingsforløb må forventes at medføre en mindre sammenlignelig patientkohorte og skabe unødig støj i forbindelse med specialets undersøgelser.

3.4.5 Behandlingsudfald

På figur 3.4 optræder behandlingsudfaldene ved tidpunkterne $t_{A,slut}$ og $t_{B,slut}$. Det første udfald er karakteriseret ved skiftet til anden behandling uanset om den tilstedeværende ordination fortsætter efter behandlingsforløbets ophør. På figuren er der tale om kortvarig polyfarmaci men det skal dog bemærkes, at der ikke skelnes mellem forskellige udfaldstyper af behandlingsudfaldet behandlingsskifte. Sådanne udfald betragtes generelt blot som negative udfald af den pågældende behandling. Figurens andet udfald er karakteriseret ved udskrivning, hvilket betragtes som et positivt udfald af den pågældende behandling jævnfør afsnit 2.2.2.

3.5 Afrunding

I dette kapitel er baggrunden for behandling med antipsykotika blevet gennemgået. Hensigten med kapitlet var at bidrage til forståelsen af grundlaget for de enkelte behandlingsforløb både med hensyn til det tilstedeværende sygdomsudtryk, de farmakologiske interventionsmuligheder og behandlingens udfald.
Det følgende kapitel vil fokusere på dokumentationen af psykosebehandlingerne i EPJ-systemet på Psykiatrisk Center Sct. Hans.

Kapitel 4

EPJ-systemet på Psykiatrisk Center Sct. Hans

4.1 Psykiatrisk Center Sct. Hans

Psykiatrisk Center Sct. Hans er geografisk placeret med sengeafdelinger i Roskilde og ambulatorium i København. For behandlinger på sengeafdelingerne er der tale om heldøgnsindlæggelser, mens ambulatoriet alene varetager den ambulante behandling; det vil sige behandlingsforløb i hospitalsregi uden indlæggelse, hvor patienten har periodisk kontakt med centrets sundhedsfaglige personale. Da en væsentlig del af behandlingen ved de ambulante forløb foregår i patientens eget hjem, i bosted eller lignende, er ambulante forløb ikke medtaget i specialets undersøgelser.

4.1.1 Specialopgaver

Centrene i Region Hovedstadens psykiatri, som Psykiatrisk Center Sct. Hans er en del af, varetager hver især en række behandlingsmæssige specialopgaver. For Psykiatrisk Center Sct. Hans' vedkommende drejer det sig om retspsykiatri, rehabilitering og dobbeltdiagnosebehandling ved misbrugsrelaterede lidelser. Centrets patienter henvises direkte fra enten egen læge, distrikspsykiatrien, anden centerafdeling eller kriminalforsorgen. Centret har således ingen akut modtagelse [Schneider, 2008b] og [Schneider, 2008a].

4.1.2 Centerstruktur

I den periode EPJ-systemet har været i brug på Psykiatrisk Center Sct. Hans, har centret undergået en række administrative ændringer senest med centerdannelsen i midten af 2007, hvor Sct. Hans overgik fra psykiatrivirksomhed til psykiatrisk center [Region H, 2007]. Og det blot et halvt år efter at *virksomheden* var ophørt med at være hospital [Schneider, 2007].
Dette og andre administrative faktorer har naturligvis haft en effekt på centrets organisation hvor blandt andet sengeafdelinger i tidens løb er blevet lukket, splittet

op eller slået sammen med andre. Centret består i dag af tre faste sengeafdelinger: afdeling L, M og R, hvor *hospitalet* for bare et årti siden bestod af afdelingerne P, K, M, R og U [Fog, 1995]. Disse afdelingsbetegnelser figurerer dog alle fortsat ved opslag i EPJ-databasen.

Afdeling L

Afdeling L råder over 118 døgnpladser fordelt på 10 afsnit, hvoraf de to er lukkede. Et afsnit betegnes som lukket, når det kun er muligt for patienten at forlade afsnittet efter aftale med en af afsnittets læger eller patientens kontaktperson. Afdelingen arbejder med rehabilitering af svært psykisk syge patienter med en diagnose inden for det skizofrene (F2) eller affektive (F3) område. Rehabiliteringen sker i forløb på 4-9 måneder med kombinationer af sociale, psykologiske og medicinske behandlinger, der har til formål at *forberede* patienten til videre ambulant behandling i eget nærmiljø [Schneider, 2008b].
For indlæggelse på afdeling L skal patienten have behov for minimum et af følgende [Schneider, 2008b]:

- medicinomlægning
- ændring af uhensigtsmæssig adfærd
- genetablering af socialt netværk
- forbedring af funktionsniveau

Afdeling M

Afdeling M råder over 130 sengepladser fordelt på 7 afsnit. Patientgruppen udgøres af patienter med en *dobbeltdiagnose*, hvilket vil sige, at patienterne har pådraget sig en psykose eller anden non-psykotisk lidelse med behov for psykiatrisk behandling som følge af eller sekundært til enten misbrug eller hjerneskade [Schneider, 2008b]. Idet hovedparten af de psykiatriske patienter enten har eller undervejs i sygdomsforløbet udvikler et misbrug, hører specialambulatoriet i København naturligt ind under afdelingens administration som eksternt afsnit. Denne afdelingstilknytning influerer dog ikke på forsøgene i specialet.
Behandlingen på afdeling M tager udgangspunkt i aftaler mellem behandler og patient om både forløb og målet med behandlingen. Psykoedukation, social færdighedstræning, kognitiv terapi og medicinsk behandling udgør i samspil afdelingens primære behandlingsformer [Schneider, 2008b].

Afdeling R

Afdeling R har 80 sengepladser fordelt på 4 lukkede og 2 åbne afsnit. Afdelingen modtager patienter med såkaldt "retspsykiatriske foranstaltninger", hvilket primært

vil sige patienter, der ved domsfældelse idømmes til mentalundersøgelse eller anbringelse. De retspsykiatriske patienter har ofte psykotiske lidelser eller lider under en alvorlig personlighedsforstyrrelse med tilfælde af særlig personfarlig kriminalitet.

Tidligere psykiatriske afdelinger

Afdeling U var en afdeling for unge skizofrene, der i 2001 havde 119 sengepladser [Sebbelov, 2001]. Kompetencerne herfra er i dag overtaget af centre med specialopgaver indenfor ungdomspsykiatri. Afdeling P varetog behandling af de psykotiske langtidssyge mens patienter på afdeling K var gruppen af „københavnske svært intergrerbare psykotiske" [Neergaard, 2005]. Kompetencerne fra afdeling P og K ligger for Psykiatrisk Center Sct. Hans' vedkommende hos centrets afdeling L.

4.1.3 Afdelingstilknytning og definitioner

Det er væsentligt at bemærke, at indlæggelser sker på afdelingsniveau. At en patient er indlagt på Psykiatrisk Center Sct. Hans er således sekundær til selve indlæggelsen på den pågældende afdeling. Da specialets undersøgelser imidlertid omfatter hele centerindlæggelsen for den enkelte patient, vil denne blive benævnt *indlæggelsen* mens indlæggelser på afdelingsniveau vil blive omtalt som *afdelingsindlæggelser*. En patient kan således have flere på hinanden følgende afdelingsindlæggelser under en og samme indlæggelse.

4.2 EPJ-systemet på Psykiatrisk Center Sct. Hans

4.2.1 Begrebsdefinition

Patientjournalens indhold er fastsat ved lov nr. 451 af 22. maj 2006 om autorisation af sundhedspersoner, der ifølge [Smith, 2006] har til formål, at:

> danne grundlag for behandling af patienten, dokumentere den udførte behandling, fungere som det nødvendige interne kommunikationsmiddel mellem det personale, der deltager i behandlingen af patienten, sikre kontinuitet i behandlingen og sikre information af patienten [Smith, 2006].

Lovkravet dækker både de papirsbaserede patientjournaler og de elektroniske patientjournaler forkortet EPJ.
Forkortelsen EPJ ses ofte anvendt i betydningen *det elektroniske patientadministrationssystem*, det vil sige et hospitalssystem til administration af afdelingernes elektroniske patientjournaler. Udover selve journaldelen vil et system også inkludere administrationen af andre hospitalsfaglige ydelser som medicinering og laboratorieprøver. Selvom man kan argumentere for, at betegnelsen elektronisk patientadministrationssystem i højere grad præciserer indholdet for et sådan system, har jeg i specialet valgt at benytte betegnelsen EPJ-systemet, da jeg mener, at betegnelsen er mere rammende

for det aktuelle system på Psykiatrisk Center Sct. Hans. Til forskel fra andre patientadministrationssystemer i regionen er samtlige funktioner i Psykiatrisk Center Sct. Hans' system velintegreret i de enkelte elektroniske patientjournaler [Deloitte, 2007]. I specialet anvendes en eksplicit navngivning til skelnen mellem de to niveauer. Derfor benyttes betegnelsen EPJ-systemet om administrationsværktøjet og elektronisk patientjournal (EPJ) om den enkelte journal.

4.2.2 Sundhedsfagligt indhold i EPJ-systemet

EPJ-systemet på Psykiatrisk Center Sct. Hans indeholder et overbliksbillede for den enkelte patients afdelingsindlæggelser, hver omfattende følgende underordnede struktur: en notatdel, en medicineringsdel samt en laboratorieundersøgelsesdel. Notatdelen omfatter de egentlige elektroniske patientjournaler, mens medicineringsdelen vedrører medicinordinationer. Laboratorieundersøgelsesdelen omfatter blandt andet rekvisitioner og undersøgelsessvar men er ikke medtaget i specialets undersøgelser, da der ikke foretages faste registreringer i denne del ved samtlige indlæggelser.
Følgende gennemgang tager udgangspunkt i EPJ-systemets interface, som sundhedspersonalet benytter ved interaktion med systemet.

Overbliksbilledet

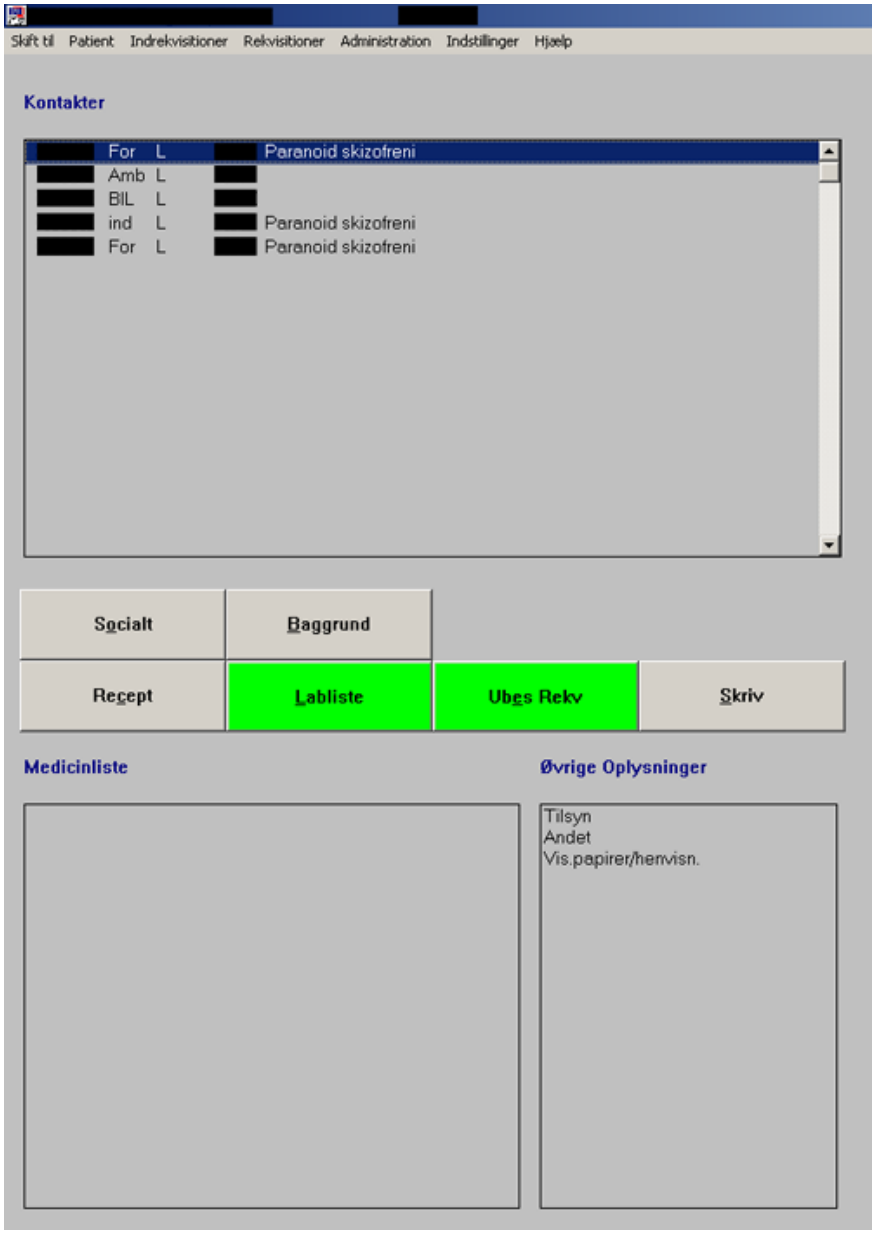
På figur 4.1 er EPJ-systemets overbliksbillede gengivet. Overbliksbilledet giver adgang til de klinisk anførte optegnelser for patientens kontakt med Psykiatrisk Center Sct. Hans. Overordnet opdeles adgangen på de forskellige undersøgelser, ambulatoriekontakter, prøvesvar for laboratorieundersøgelser og afdelingsindlæggelser. Ved afdelingsindlæggelser og generelle undersøgelser er overbliksbilledet desuden tilføjet angivelse af patientens hoveddiagnose ligesom indlæggelsesforløb desuden er suppleret med angivelse af aktive præparatordinationer.

Notatdelen

Figur 4.2 illustrerer fremstillingen af oplysninger i notatdelen. Figuren illustrerer en såkaldt epikrise, det vil sige et klinisk tilbageblik over et patientforløbet ved udskrivelsen. Dette anonymiserede eksempel er en patient fra afdeling R. Udover epikrisen indeholder notatdelen for hver afdelingsindlæggelse også indlæggelsesnotater.

Medicineringsdelen

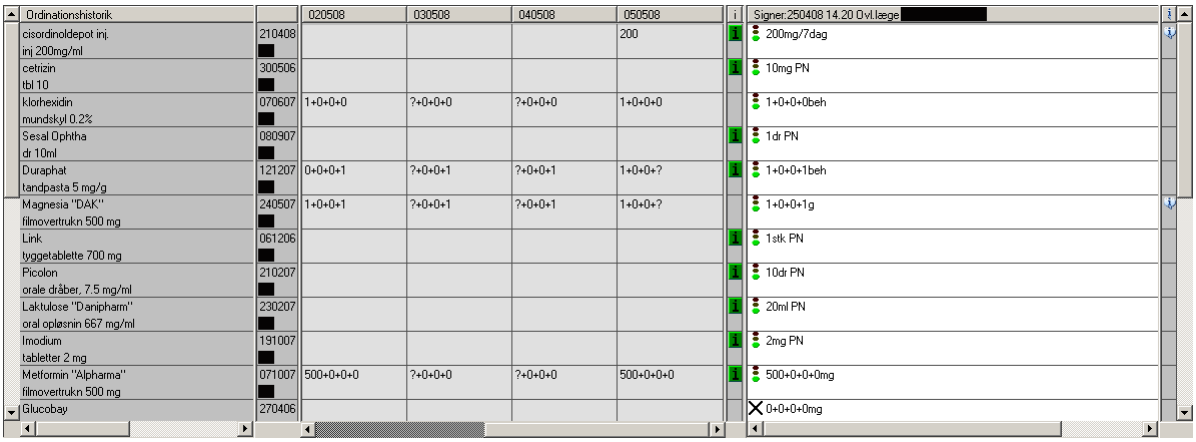
Medicineringsdelen omfatter samtlige ordinationer for hele afdelingsindlæggelsen. Figur 4.3 illustrerer medicinskemaet for en af centrets patienter. Første kolonne (ordinationshistorik) angiver præparatnavn og indgivelse. Den anden kolonne angiver startdatoen for den enkelte ordination og tredje kolonne angiver uddelinger på hver kalenderdato. Fjerde kolonne angiver tilstedeværelse af yderligere information om ordinationen mens femte og sidste kolonne benyttes om dosering, uddelingsfrekvens og



Figur 4.1: Overbliksbillede for patient på afdeling L

Figur 4.2: Udskrivningsnotat (epikrise) for patient på afdeling R

andre særlige forhold ved ordinationen såsom midlertidig ophør eller behovsmedicinering.



Figur 4.3: Medicinskema i EPJ-systemet

4.3 EPJ-systemets database

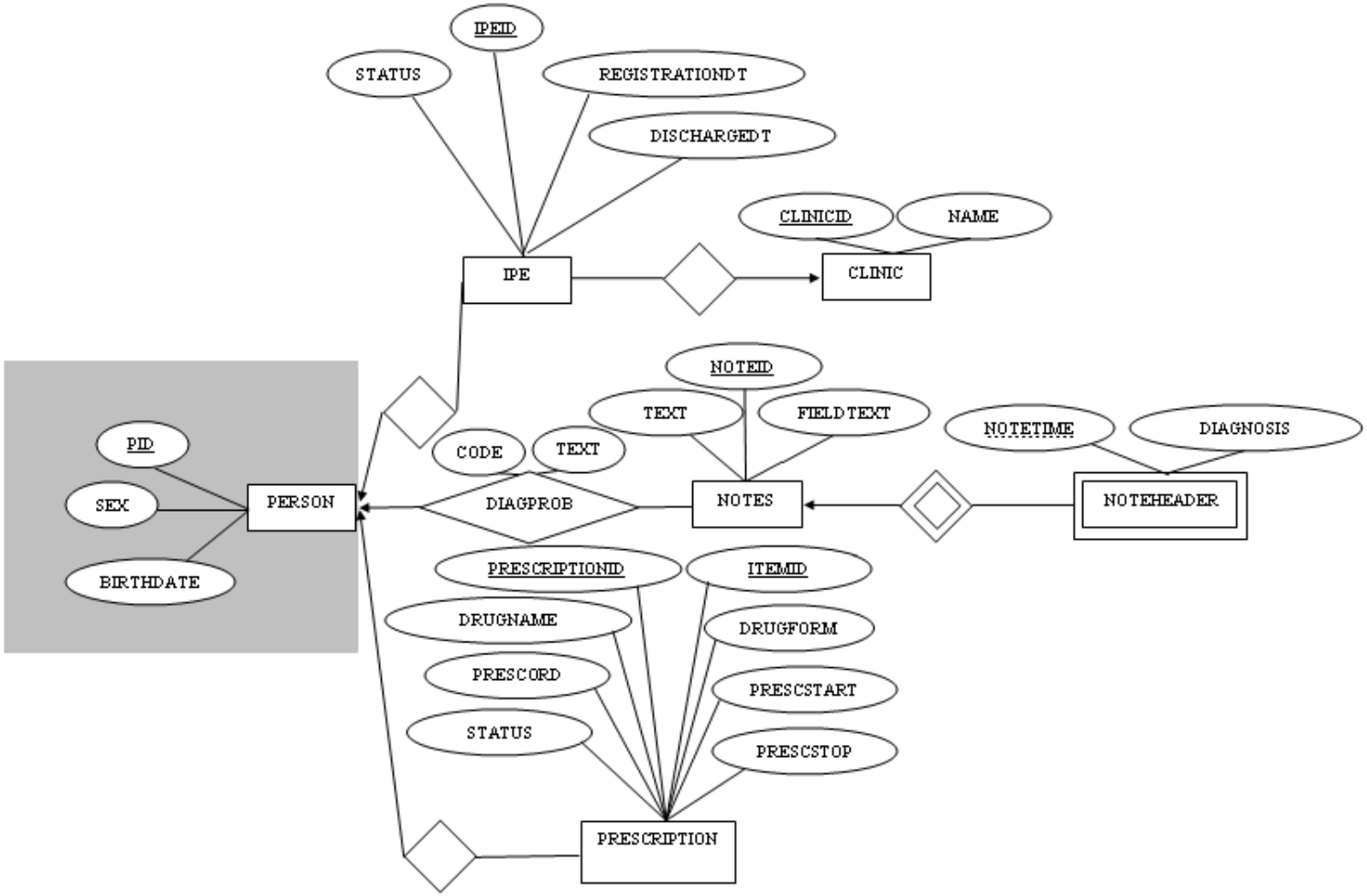
4.3.1 Baggrund

EPJ-systemet anvender IBMs DB2, der er en relationel database. Det betyder at informationerne i databasen er gemt i tabeller og at relationerne mellem disse tabeller udgør strukturen i databasen. Idet kun et udsnit af databasens tabeller, vedrører patientbehandlingen har jeg udarbejdet et E/R-diagram[1] over denne struktur. Diagrammet er vist på figur 4.4.
Diagrammets entiteter er givet ved de udvalgte tabeller i databasen. Tabellerne og de felter der er indeholdt i hver tabel i databasen er kort gennemgået i tabel 4.1.

4.3.2 Struktur og konventioner

Selvom EPJ-systemet er begrundet i et utal af formålserklæringer, er det væsentlige fokus - med rette - lagt på dokumentation, fleksibilitet og overskuelighed. Dette er muligvis en væsentlig årsag til, at et gennemgående træk ved registreringerne i EPJ-systemet er at særligt datofelter er præget af fejl. Simple konventioner introduceret på registreringssiden kunne formegentlig have reduceret fejlprocenten væsentligt. Som eksempler på sådanne konventioner kan nævnes et krav om, at en given udskrivningsdato ikke kan være registreret før en indlæggelsesdato eller at systemets registrering af patientens fødselsdato baseres på det indtastede cpr-nummer - hvis et sådan findes - for at undgå uoverensstemmelser mellem de to.

[1] For en gennemgang af opbygningen af E/R-diagrammer og notationer henvises til [Silberschatz *et al.*, 2006] side 214.



Figur 4.4: E/R-diagram for det udsnit af databasen der anvendes til specialets undersøgelser. Den grå markering omkring tabellen PERSON indikerer at der er tale om udsnittets mest centrale tabel.

PERSON	
PID	Løbenummer for patienten
SEX	Køn
BIRTHDATE	Fødselsdato

IPE	
IPEID	Løbenummer for afdelingsindlæggelsen
STATUS	Aktivstatus for indlæggelsen (aktiv, afsluttet eller annulleret)
REGISTRATIONDT	Indskrivningsdato
DISCHARGEDT	Udskrivningsdato

CLINIC	
CLINICID	Løbenummer for afdelingen
NAME	Afdelingsnavn

DIAGPROB	
CODE	Diagnosekode
NAME	Diagnosenavn[1]

NOTES	
NOTEID	Løbenummer for journalnotatet
FIELDTEXT	Løbenummer for notatfeltet
TEXT	Journaltekst

NOTEHEADER	
NOTETIME	Dato for journaloptegnelsen
DIAGNOSIS	Diagnose[2]

PRESCRIPTION	
PRESCRIPTIONID	Løbenummer for ordinationen
ITEMID	Løbenummer for uddelingen
DRUGNAME	Præparatnavn
DRUGFORM	Præparatform (tablet, kapsel osv.)
PRESCORD	Ordinationstype (fast eller ved behov)
STATUS	Aktivstatus for ordinationen (aktiv, afsluttet eller annulleret)
PRESCSTART	Startdato for ordinationen
PRESCSTOP	Ophørsdato for ordinationen

Tabel 4.1: Anvendte tabeller fra EPJ-systemets database med udgangspunkt i E/R-diagrammet på figur 4.4. De enkelte tabeller er angivet med deres tabelnavn (skrevet med fed) og med en kort beskrivelse af de indeholdte felter. ([1]) angiver at der er tale om en kategoriseret diagnose, det vil sige at diagnosen er angivet som hoved- eller bidiagnose. ([2]) angiver at diagnosen ikke er kategoriseret som hoved- eller bidiagnose.

Udover fraværet af konventioner er en anden væsentlig udfordring til sikring af valide data fra EPJ-systemet, at systemet er baseret på en generel systemmodel fra systemets udbyder, der har udviklet sig over tid. Således optræder der et løbende skifte i omfanget af registreringer og anvendelsen af databasens tabeller.
Problemer som fraværet af konventioner og inkonsistens i registreringsmåden har lille eller ingen indflydelse på dagligdagsbrugen af materialet i den enkelte EPJ. Til gengæld er indflydelsen stor, når det kommer til automatiseret indsamling og analyse af større datamængder, hvilket vil indgå som et væsentligt og nødvendigt tema i de følgende kapitler.

4.4 Indhold i patientjournalen

Indholdet i patientjournalen er baseret på såkaldte *journalnotater*. I specialet anvendes betegnelsen journalnotat om samtlige notater der registreres i EPJ-systemet, selvom notaterne navngives i klinikken efter deres formål. I tabel 4.2 er angivet en række hyppigt anvendte notater med angivelse af notatnavn og sundhedsfagligt indhold. Årsagen til at der ikke anvendes samme navngivning i specialet skyldes at der alene skelnes mellem *samtlige* notattyper og notater udarbejdet af læger. For sidstnævntes vedkommende drejer det sig om notater angivet med * i tabel 4.2.
I databasen for EPJ-systemet er det tabellen NOTES, der anvendes til registrering af de enkelte journalnotater.

4.5 Afrunding

Kapitlet havde til formål at klarlægge hvordan patientinformationerne er registreret i EPJ-systemet på Psykiatrisk Center Sct. Hans.
I det følgende kapitel vil data mining blive gennemgået med fokus på videnuddragning fra databaser med særligt fokus på kliniske data, som kendetegnet ved data registreret i EPJ-systemet.

Notatnavn	Sundhedsfagligt indhold
Forundersøgelse*	Notatet udfærdiges inden indlæggelsen med angivelse af blandt andet: • Indlæggelsesårsag • Tidligere psykiatrisk indlæggelse • Medicinanamnese (aktuel medicinstatus) • Symptombillede • Misbrug • Diagnose (hoved- og bidiagnose)
Indlæggelsesnotat*	Indeholder samme oplysninger som notatet *forundersøgelse* dog eventuelt suppleret med tilkomne oplysninger eller laboratoriebestillinger (eksempelvis måling af lipider i blodet).
Modtagelse	Gennemgang af foranstaltninger eller særlige forhold ved indlæggelsen oplyst til patienten.
Indlæggelsessamtale(*)	Samtale med patienten om den forestående behandling.
Rehabiliteringsplan(*)	Rehabiliterende initiativer ved indlæggelsen.
Behandlingsplan*	Vedrører primært de farmakologiske interventioner.
Plejeplaner	Planlægning af pleje forbundet med indlæggelsen.
Medicinafvigelser*	Afvigelser fra den planlagte medicinske behandling med angivelse af årsag.
Sygeplejestatus	Status for pleje af patienten med angivelse af patientens respons på behandlingen.
Behandlingsnotat*	Oplysninger forbundet med den farmakologiske del af behandlingen og symptombilledet.
Sygeplejenotat	Oplysninger forbundet patientens pleje.
Særlige aftaler(*)	Særlige aftaler mellem afdelingen og patienten.
Andre notattyper	Gruppeterapi / ergoterapeut / socialrådgiver / erhvervsterapi / psykolog.
Udskrivelsesnotat*	Status for behandlingen ved udskrivning.
Epikrise(*)	Indeholder samme oplysninger som notatet *udskrivelsesnotat* suppleret med oplysninger om plan for udskrivelse af patienten.

Tabel 4.2: Anvendte notatnavne for journalnotater registreret i EPJ-systemet. * angiver at notatet alene udfærdiges af læger mens (*) angiver at både læger og andre personalegrupper bidrager med oplysninger til notatet. Øvrige notater suppleres normalt ikke med oplysninger fra afdelingens læger.

Kapitel 5

Data mining

Afsnittet har til hensigt at introducere begrebet data mining i relation til specialets undersøgelser samt at give en indføring i de valgte metoder, som finder anvendelse i specialets eksperimenter.

5.1 Begrebsdefinition

Begrebet *data mining* henviser til udvinding af data som parallel til udvinding af råstoffer fra undergrunden og benyttes om den proces at uddrage hidtil ukendte sammenhænge og mønstre fra større datamængder [Prather et al., 1997].

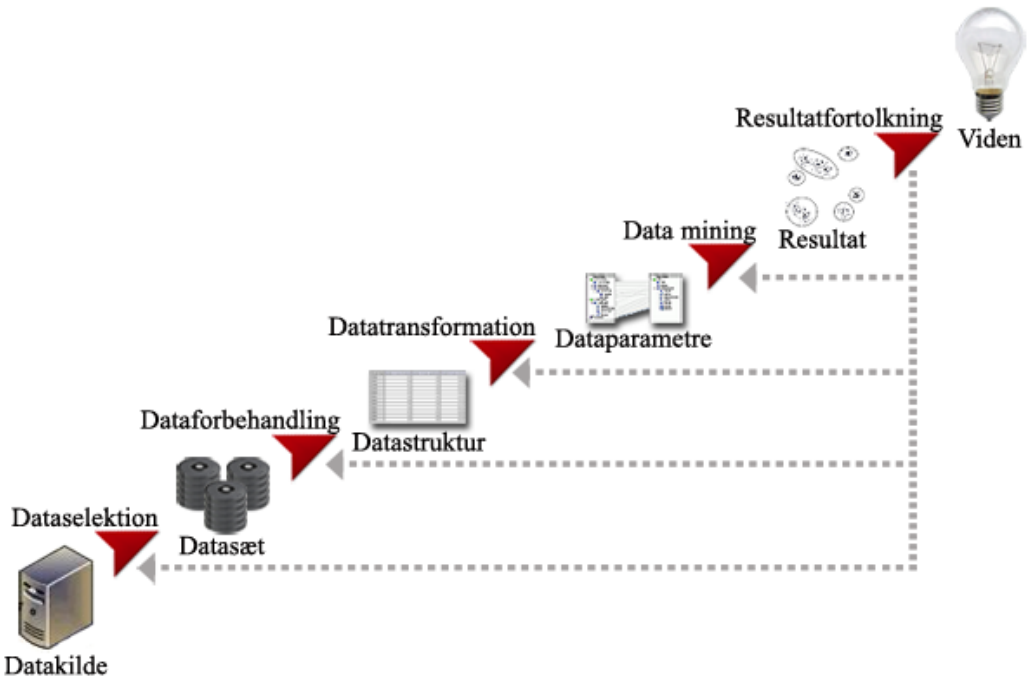
I litteraturen anvendes begrebet data mining på to niveauer: et overordnet niveau, hvor begrebet benyttes om hele *processen* fra datakildens klargøring til det endelige analyseresultat foreligger; samt et *metode*-niveau, hvor begrebet dækker over den valgte data mining metode til udredning af sammenhænge og strukturer. I sidstnævnte tilfælde udgør data mining metoden blot et enkelt trin i den samlede proces. Jeg vil i specialet sondre mellem de to niveauer med reference til enten processen eller metoden[1].

Tekst data mining er en specialisering af data mining-begrebet, som vedrører de metoder, der knytter sig til anvendelsen af tekstdokumenter som datakilde, hvor uddragningen af sammenhænge metodisk set er fokuseret på ustruktureret tekst [Razvan and Bunescu, 2005]. I specialet anvendes data mining og tekst data mining som synonyme begreber, hvor specialiseringen i det enkelte tilfælde afgøres af det valgte *datasæt*. Datasættet udgøres af de elementer (eller dokumenter ved tekst data mining), som skal analyseres.

[1] I den engelsksprogede litteratur anvendes både *Knowledge Discovery in Databases* og *data mining* som betegnelser for det overordnede procesniveau. Idet ingen af begreberne foreligger i en fastlagt dansk form har jeg valgt den beskrevne begrebslige niveaumæssige skelnen.

5.2 Data mining processen

Data mining processen er en trinvis proces, som udover den valgte data mining metode omfatter dataselektion, dataforbehandling, datatransformation og resultatfortolkning. [Fayyad *et al.*, 1996] har foreslået en model over forløbet, som vil blive anvendt til beskrivelse af processen. Modellen er illustreret på figur 5.1. Det er væsentligt at pointere, at modellen er et generaliseret billede af processen. Nogle proces-former vil for eksempel bytte om på trinene eller inddrage elementer fra samme trin i et og samme trin, men modellen er valgt, fordi den giver et godt overblik over processens delelementer, der uanset konfigurationen er nødvendige for at udføre succesfuld data mining. Modellen fremtræder umiddelbart statisk i den henseende, at hvert trin leder frem til det følgende i processen, men det er væsentligt at bemærke, at forløbet er dynamisk, idet parametre for de enkelte trin i processen kan justeres undervejs (som angivet ved de stiplede pile), hvorved forskellige udfald kan sammenlignes.



Figur 5.1: Model over data mining processen frit efter [Fayyad *et al.*, 1996]. Processen tager udgangspunkt i dataselektion fra en given datakilde og derpå en forbehandling af de selekterede data, der fører til en strukturering af disse data. Ved datatransformationen identificeres specifikke parametre i de strukturerede data som efterfølgende benyttes i den valgte data mining metode. Resultatet herfra underkastes slutteligt en fortolkning. På et hvilket som helst tidspunkt er det muligt at træde et eller flere trin tilbage i processen og genoptage undersøgelsen fra dette trin, hvilket er illustreret ved de stiplede pile.

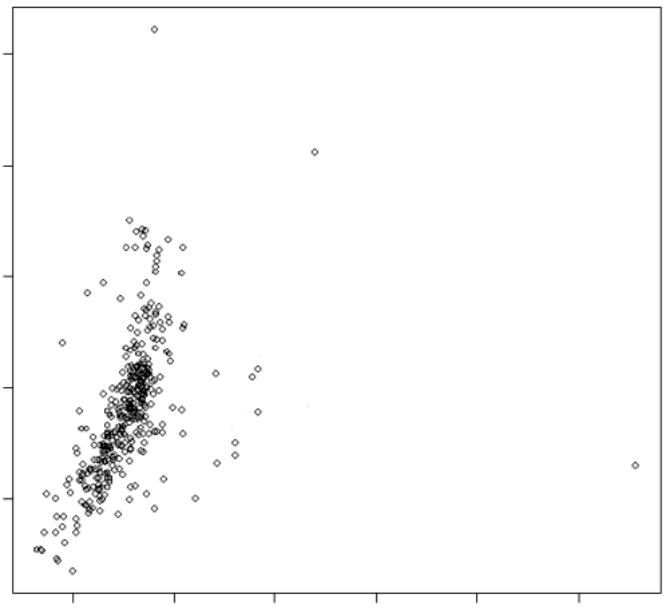
5.2.1 Dataselektion

Dataselektion beskriver de metoder, der knytter sig til *udvælgelsen* af data fra en eller flere datakilder og er illustreret på figur 5.1 som processens første trin. En

datakilde vil som udgangspunkt, og som det er tilfældet i specialet, udgøres af en database. Udover en bestemmelse af datakilden vil dataselektion indbefatte det udsnit af datakilden, der ønskes undersøgt. Udvælgelsen af dette udsnit vil være styret af undersøgelsens formål og kan udgøres af enten hele eller en delmængde af de tilgængelige data. Udsnittet kan fastlægges ud fra strukturen i databasen, et fokus på bestemte datatyper eller ganske enkelt som en simpel stikprøve [Fayyad *et al.*, 1996].
På baggrund af dataselektionen fastlægges datasættet for den pågældende undersøgelse.

5.2.2 Dataforbehandling

Efter at have fastlagt et datasæt for undersøgelsen, vil det ofte være nødvendigt at fjerne *støj*. Støj repræsenterer i denne sammenhæng manglende værdier, duplikater og inkonsistens mellem de registrerede data [Fayyad *et al.*, 1996]. Inkonsistens kan komme til udtryk på mange måder afhængig af årsagen. Sidstnævnte skyldes ofte den tidsmæssige dimension som for eksempel spredte opdateringer over en længere årrække, ændring i brugsmønstre, administrative omlægninger osv.
Forestiller man sig en simpel repræsentation af sit datasæt i et 2-dimensionelt rum, vil det generelt være de punkter, som i deres placering i høj grad afviger fra de øvrige punkter; de såkaldte *outliers*, der er udtryk for datasættets støj. Dette er eksemplificeret i figur 5.2, hvor særligt punktet placeret i figurens højre side afviger væsentligt fra de øvrige.



Figur 5.2: Punkter i et 2-dimensionelt rum med en sky af punkter i nederste venstre kvadrant omgivet af tre outliers.

Hensigten med dataforbehandling eller *data preprocessing* er at minimere effekten af

støjen på det valgte datasæt. Den anvendte metode til dataforbehandling vil dog i høj grad være dikteret af målet med undersøgelsen [Fayyad *et al.*, 1996]. Det skal slutteligt bemærkes, at de skitserede eksempler er relativt simple. Generelt kræves således et særdeles godt kendskab til datasættet, før det er muligt at fastlægge et mål for fjernelse af støjen.
Begreber som *aggregering*, *generalisering*, *normalisering* og *konstruktion* knytter sig til en række metoder, som finder anvendelse i denne sammenhæng [Han and Kamber, 2001].
For at eksemplificere disse begreber vil den følgende gennemgang tage udgangspunkt i tabel 5.1, der er et eksempel på en tabel med kolonnerne *patient*, *præparat* og *dosis*. Den valgte tabel tjener alene til det illustrative formål at eksemplificere de anvendte begreber og er således *ikke* udtryk for den valgte data mining metode til dataforbehandling af EPJ-systemets database. For en sådan gennemgang henvises til afsnit 5.4.4.

Patient	**Præparat**	**dosis**
Gert	Abilify	6 mg
Gert	Nozinan	50 mg
Gert	Nozinan	400 mg
Ulla	Leponex	450 mg
Tom	Zyprexa	20 mg
Tom	Zyprexa	20 mg
Kim	Clozapin	900 mg

Tabel 5.1: En eksemplificeret tabel med dosisbestemmelser for patienterne Gert, Ulla, Tom og Kim.

Aggregering

Aggregering henviser til sumering af værdier, hvorved aggregering kan danne grundlag for et mere forsimplet og velstruktureret datasæt.
Et blik på tabel 5.1 afslører, at patienterne Gert og Tom begge har dobbelt forekomst af ordinationer for henholdsvis Nozinan og Zyprexa. Aggregering af disse forekomster vil føre til tabel 5.2, hvor netop de fremhævede felter er fremkommet ved aggregering[2].

Generalisering

Generalisering henviser til en konceptualisering eller overordnet repræsentation af de selekterede data. Et eksempel herpå kunne være at erstatte præparatangivelser med klassifikation af præparatet som henholdsvis typisk eller atypisk antipsykotika.
Tabel 5.3 illustrerer generalisering af præparatnavnene med udgangspunkt i den aggregerede dosistabel (tabel 5.2).

[2]En sådan sumering af dosisstørrelser vil i en virkelig tabel over præparatordinationer naturligvis være diskutabel fra et medicinalbiologisk synspunkt og anvendes da også kun med det formål at illustrere aggregeringen på nummeriske værdier.

Patient	Præparat	Dosis
Gert	Abilify	6 mg
Gert	Nozinan	450 mg
Ulla	Leponex	450 mg
Tom	Zyprexa	40 mg
Kim	Clozapin	900 mg

Tabel 5.2: Dosisbestemmelser hvor felter med ordinationer af Nozinan og Zyprexa fra tabel 5.1 er fremkommet ved aggregering.

Patient	Præparat	Dosis
Gert	Atypisk	6 mg
Gert	Typisk	450 mg
Ulla	Atypisk	450 mg
Tom	Atypisk	40 mg
Kim	Typisk	900 mg

Tabel 5.3: Tabel med generaliseret præparatangivelse.

Normalisering

Normalisering indenfor data mining[3] henviser til *skalering* af værdier i datasættet. Eksempelvis kan dosisstørrelser for patienternes præparatordinationer opstilles procentuelt i forhold til den højest anvendte daglige dosis i datasættet. Hvis højeste dosisforekomst for Nozinan er 600 mg, vil dosisforekomster ved denne værdi efter normalisering fremstå som 1, mens Nozinandoser på 450 mg dagligt vil fremstå som $(\frac{450}{600}) = 0{,}75$.
Tabel 5.4 illustrerer eksemplet med normalisering af dosisstørrelser.

Patient	Præparat	Dosis
Gert	Abilify	0,2
Gert	Nozinan	0,75
Ulla	Leponex	0,5
Tom	Zyprexa	1
Kim	Clozapin	1

Tabel 5.4: Tabel med dosisbestemmelser hvor dosisstørrelserne (fremhævet) er normaliseret.

Konstruktion

Konstruktion refererer til konstruktion af supplerende tabeller eller attributter på baggrund af den eksisterende struktur til brug for data mining-metoden.

[3]Anvendelse af begrebet normalisering i data mining-sammenhæng er væsentlig forskellig fra den mere anvendte betydning ved omtale af databaser , der vedrører metoder til relationsmæssig database-strukturering

Et eksempel på en sådan konstruktion kan være tilføjelsen af en attribut med angivelse af præparatets indholdsstof. Tabel 5.5 illustrerer dette.

Patient	Præparat	Dosis	Indholdsstof
Gert	Abilify	6 mg	Aripiprazol
Gert	Nozinan	450 mg	Levomepromazin
Ulla	Leponex	450 mg	Clozapin
Tom	Zyprexa	40 mg	Olanzapin
Kim	Clozapin	900 mg	Clozapin

Tabel 5.5: Tabel med dosisbestemmelser tilføjet kolonne (fremhævet) med angivelse af præparatets indholdsstof.

De omtalte metoder kan benyttes uafhængigt af hinanden og i et sammenspil langt mere komplekst end i de beskrevne eksempler.

5.2.3 Datatransformation

Datatransformation er en klargøring af datasættet til brug for den valgte data mining metode. Ved datatransformationen fastsættes et *similaritetsmål*, der er bestemmende for ligheden mellem datasættets elementer. Oftest udtrykkes ligheden eller similariteten ved et tal i intervallet [0,1], hvor 0 angiver det absolutte fravær af lighed, 1 den absolutte lighed og værdierne herimellem udtrykker den procentuelle lighed. I forbindelse med tekst data mining processen, benyttes en række mere statistisk funderede mål for ligheden (se afsnit 5.4.5), mens ligheden mellem to nummeriske værdier kan udtrykkes direkte procentuelt, som den laveste værdis procentdel af den højeste.

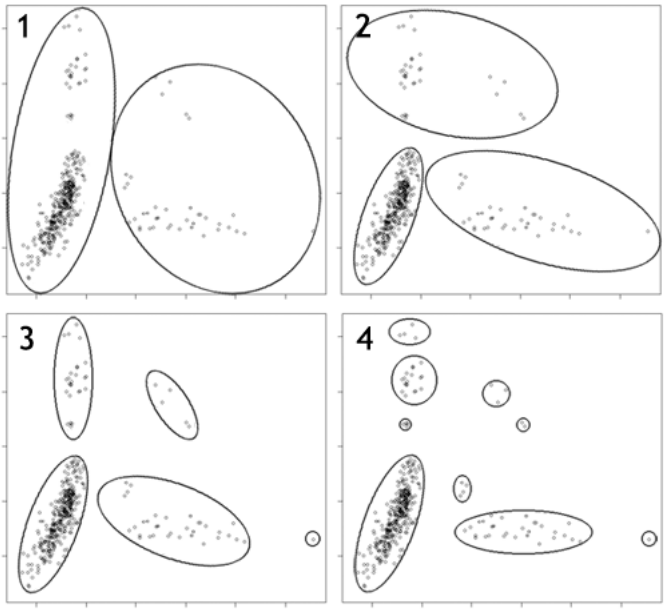
5.2.4 Data mining som metode

En længere række data mining metoder findes beskrevet i litteraturen. Overordnet set kan metoderne ordnes efter den pågældende *intention* der knytter sig til metoderne. Intentionen for de eksisterende data mining metoder er enten *deskriptiv* eller *prædiktiv*. De deskriptive metoder retter sig mod at uddrage hidtil ukendte træk, imens de prædiktive metoder retter sig mod at bestemme en model, der på baggrund af træningssæt bliver i stand til at klassificere fremtidige datasæt [Han and Kamber, 2001]. De deskriptive metoder karakteriseres således som *usuperviserede*, idet klassifikationen foretages på baggrund at eller flere similaritetsmål, imens de prædiktive metoder karakteriseres som *superviserede* qua den iterative læringsproces.

De hyppigst anvendte data mining metoder er *clustering*, *associationsanalyse* og *decision trees*, hvor de to første er deskriptive metoder mens den sidste er en prædiktiv data mining metode.

Clustering

Clustering omtales også som clusteranalyse og refererer til opdelingen af et datasæts elementer i 2 eller flere grupper eller *clusters* med det formål, at hvert element i det enkelte cluster har større *lighed* med elementer fra samme cluster end med elementer fra alle øvrige clusters [Han and Kamber, 2001].
Ligheden bestemmes matematisk ud fra det valgte similaritetsmål, der benyttes til udregning af en afstand mellem alle datasættets elementer, hvilket clusteranalysen baserer sin opdeling på.
Figur 5.3 illustrerer fire forskellige clusteropdelinger af *samme* datasæt. Figur 5.3 (1) illusterer opdelingen i to clusters, (2) opdelingen i tre clusters, (3) opdelingen i fem clusters og (4) opdelingen i 9 clusters. Som figuren illustrerer, er clusterdelingerne forbundet med spredningen i punkternes afstand. Således er det clusteret med den største spredning, der vil være mål for konsekutive clusterdelinger. Det højre cluster på figur 5.3 (1) fremstår med størst spredning og splittes i to ved datasættets opdeling i tre clusters (2).
Som figur 5.3 (4) illustrerer, vil outliers hurtigt blive placeret i mindre clusters, og clusteranalyse benyttes da også til støjreduktion i forbindelse med dataforbehandling i starten af data mining processen [Fayyad *et al.*, 1996].

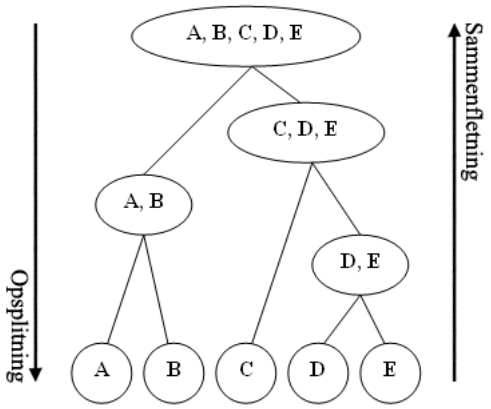


Figur 5.3: Fire forskellige clusteropdelinger af samme datasæt. (1) illustrerer opdelingen i 2 clusters, (2) illustrerer opdelingen i 3 clusters, (3) illustrerer opdelingen i 5 clusters mens (4) illustrerer opdelingen i 9 clusters.

Clustermetoder opdeles i henholdsvis hierarkisk og partitions-clustering:

Hierarkisk clustering. Hierarkisk clustering er illustreret på figur 5.4 som en træstruktur, hvor roden består af alle elementerne samlet i ét cluster, mens træets

blade udgøres af elementerne i hvert sit cluster [Han and Kamber, 2001]. Hierarkisk clustering kan enten udføres ved sammenfletning, hvor man bevæger sig fra bladene og op mod træets rod eller ved opsplitning, hvor man starter i roden og bevæger sig nedpad. Hvert trin repræsenterer således enten opsplitningen i to nye clusters eller sammenfletning af to clusters i et.

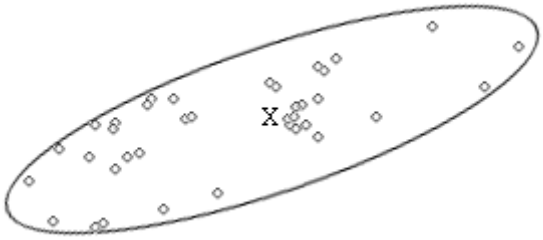


Figur 5.4: Hierarkisk clustering med opsplitning i mindre clusters når man bevæger sig nedad og sammenfletning når man bevæger sig opad. Frit efter [Han and Kamber, 2001].

Partitions-clustering. De „klassiske" clustering metoder k-means og k-medoide tilhører partitions-clustering-metoder. Til forskel fra hierarkisk clustering tager partitions-clustering udgangspunkt i et på forhånd fastsat antal clusters, hvor k'et fra førnævnte metoder begge referer til det antal clusters, som clusteranalysen skal munde ud i [Han and Kamber, 2001]. Udgangspunktet for denne metode er en opsplitning af hele datasættet i k-clusters fra begyndelsen. I de efterfølgende trin byttes elementerne ud med hinanden indtil den bedst mulige opdeling på baggrund af afstandsmålet er nået. Forskellen på k-means og k-medoide er fastlæggelsen af den middelværdi, som det enkelte cluster bygges op omkring. Ved k-means clustering er det clusterets nummeriske middelværdi, der angiver clusterets centrum (ligesom x'et på figur 5.5), imens k-medoide clustering anvender det mest centralt placerede element af clusterets elementer, som clusterets midte [Han and Kamber, 2001].
Da k-medoide anvendes som data mining metode i specialets undersøgelser, er denne algoritme gennemgået i afsnit 5.3.

Associationsanalyse

Associationsanalyse anvendes til bestemmelse af associationer eller sammenhænge i datasættet. Metoden munder ud i en fastlæggelse af et regelsæt for datasættet, der beskriver sammenhænge mellem de enkelte elementer i datassættet, eksempelvis hvilke elementer der optræder sammen [Han and Kamber, 2001]. En associationsanalyse på datasættet fra EPJ-systemet kunne eksempelvis dokumentere, at en særlig



Figur 5.5: Et cluster bestående af flere elementer centreret omkring middelværdien (markeret med x).

bivirkning optræder hyppigst sammen med et specifikt antipsykotikum.

Decision trees

Decision trees eller beslutningstræer er en metode, der er i stand til at oprette et regelsæt på baggrund af et givent datasæt, og ud fra dette regelsæt er i stand til at klassificere alle fremtidige datasæt [Han and Kamber, 2001]. Et regelsæt fra EPJ-systemet kunne eksempelvis fastlægges for psykotiske patienter. Dette regelsæt vil da kunne bruges automatiseret til at hjælpe klinikeren med at træffe beslutninger om valg af medicinsk behandling.

5.2.5 Resultatfortolkning

Selvom den valgte data mining metode fremkommer med et resultat, skal dette resultat først fortolkes for at kunne anvendes.

5.3 *k*-medoide clusteralgoritmen

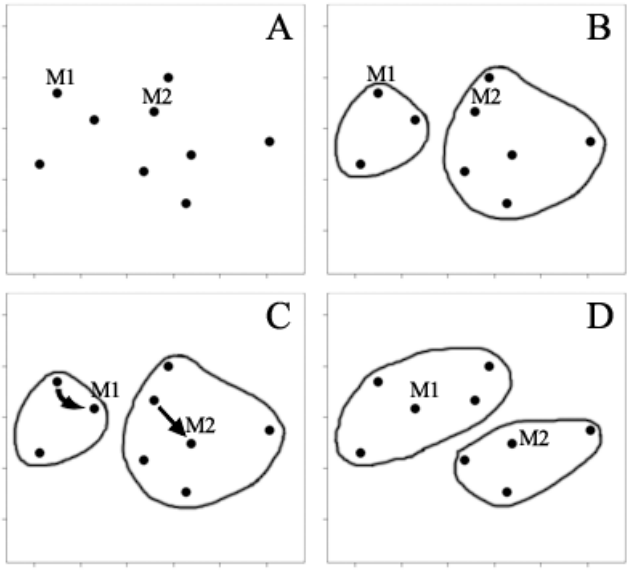
k-medoide clusteralgoritmen søger på baggrund af en forudbestemt værdi for k at opdele datasættet i k clusters. Dette sker ved trinvis tildeling af datasættets elementer i en clusterfiguration og en efterfølgende kontrol af, om afstanden mellem elementerne i clusterfigurationen er den kortest mulige. Afstandende i det enkelte cluster er afhængig af bestemmelse af det enkelte clusters mest centrale element betegnet *medoiden*, idet dette element ikke nødvendigvis er identisk med clusterets *middelværdi*.
Tildeling af medoide-status til et element i datasættet sker indledningsvist tilfældigt. Dette påvirker dog ikke det endelige udfald, idet trinene i algoritmen fører til en ensartet clusteropdeling uanset udgangspunktet. De enkelte trin i algoritmen er præciseret i tabel 5.6 på baggrund af fremstillingen af [Han and Kamber, 2001].
Figur 5.6 illustrerer algoritmens funktion ved opdeling i 2 clusters. På figurens billede A tildeles to elementer tilfældigt medoide-status (trin 1 og 2 fra tabel 5.6). På figurens billede B tilføjes de øvrige elementer det cluster, hvor elementerne er placeret nærmest clusterets medoide (trin 3 og 4 fra tabel 5.6). Ved efterfølgende beregning identificeres de elementer med laveste afstand til de øvrige (trin 5 og 9 fra tabel 5.6), og disse

k-medoide clusteralgoritme
Input: Bestem antal clusters k og tilføj et datasæt
Output: k clusters med datasættets elementer

1:	Placér tilfældigt k elementer fra datasættet enkeltvis i hvert cluster.
2:	Elementet i hvert cluster noteres som **medoide** for sit respektive cluster.
3:	Gennemløb hver af datasættets øvrige elementer.
4:	Hvert element tilføjes det cluster, hvor elementets afstand til clusterets medoide er kortest.
5:	Når alle elementerne er gennemløbet, udregnes afstanden mellem medoiderne og elementerne i det enkelte cluster.
6:	Summen af alle afstande noteres.
7:	Ved konsekutive kørsler sammenlignes summen af alle afstande med de tidligere kørsler: **HVIS** summen er konstant **AFSLUT** **ELLERS FORTSÆT:**
8:	Identificér det element i hvert cluster, der har lavest afstand til clusterets øvrige elementer.
9:	Dette element beholdes i clusteret, imens de øvrige fjernes.
10:	Spring til punkt **2** og gentag kørslen.

Tabel 5.6: Delelementer i k-medoide clusteralgoritmen [Han and Kamber, 2001].

tildeles nu medoide-status som illustreret på figurens billede C. Ved efterfølgende kørsel tildeles de øvrige elementer, og to nye clustre dannes med kortest afstand mellem elementerne.



Figur 5.6: Illustration af *k*-medoide clusteralgoritmens clustering. I (A) tildeles tilfældigt medoide-status til to elementer (M1 og M2) og der dannes to grupper af elementer ved tildeling af de resterende elementer til den gruppe, hvor afstanden fra elementet til M1 eller M2 er mindst (B). Efterfølgende udvælges det mest centrale element i hver gruppe (C), der nu kommer til at udgøre medoiderne for de to grupper (D).

5.4 Tekst data mining

I den relationelle database lagres informationerne i tabellernes felter. Indholdet i disse felter kan noget forenklet udtrykkes som enten en numerisk værdi eller en tekststreng. Ved sammenligning af simple numeriske værdier som eksempelvis løbenumre for patienter (udtrykt ved feltet PID i tabellen PERSON fra EPJ-systemets database[4]) er det relativt enkelt at afgøre, hvor tæt løbenumrene er placeret ved brug af en simpel intervalakse afgrænset af højeste og laveste værdi for de pågældende løbenumre. Er der derimod tale om en sammenligning af tekststrenge, der hver udgøres af et journalnotat (udtrykt ved feltet TEXT i tabellen NOTES i EPJ-systemets database[5]), stiger kompleksiteten i opgaven betydeligt. Med tekst data mining medfølger imidlertid et metodeapparat til at håndtere en sådan opgave. Processen fremstår identisk med det hidtil gennemgåede, imens det metodiske niveau kræver enkelte udredninger.

[4]Se tabel 4.1.
[5]Se tabel 4.1.

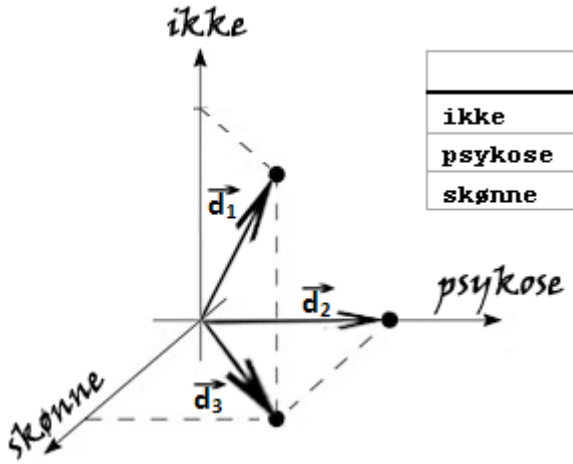
5.4.1 Dokumentbegrebet

Ved tekst data mining siges datasættet D at udgøres af *dokumenter*. Et dokument d_i skal i denne sammenhæng ses som en afgrænset mængde af ord benævnt *termer*. Vokabulariet for dokumentet d_i udgøres af samtlige termer T i d_i og kan således opstilles på formen angivet i 5.1 [Buckley , 1993].

$$d_i = (t_1, t_2, ..., t_n) \tag{5.1}$$

5.4.2 *Vector space* modellen

En hyppigt anvendt metode til at fastlægge similariteten mellem dokumenter er at udtrykke det enkelte dokument som en *vektor* [Berry *et al.*, 1999]. Der er tale om en matematisk abstraktion med udgangspunkt i *vector space modellen* [Salton, 1968], hvor vektorrummets akser udgøres af det samlede vokabularium for de dokumenter, der skal sammenlignes, hvilket er eksemplificeret på figur 5.7. Som notation for en vektor benyttes betegnelsen $\overrightarrow{v}$ hvorfor dokumentvektoren for d_i betegnes $\overrightarrow{d_i}$. Fordelen ved at repræsentere dokumenterne som vektorer er at similariteten mellem dokumenterne kan beskrives ved hjælp af et geometrisk mål.
Vektorrummets dimensioner er givet ved antallet af termer i vokabulariet og betegnes som det n-dimensionelle vektorrum, hvor n er antallet af termer i vokabulariet. På figur 5.7 er dokumentvektorerne $\overrightarrow{d_1}$, $\overrightarrow{d_2}$ og $\overrightarrow{d_3}$ fremstillet i et 3-dimensionelt vektorrum med udgangspunkt i dokumenterne d_1, d_2 og d_3. Da det ikke er muligt at repræsentere en figur i mere end tre dimensioner, er figurens vokabularium kun baseret på de tre termer „*ikke*", „*psykose*" og „*skønne*".
Ved at udtrykke vokabulariet for dokumenterne ved hjælp af et matrice, kan dette benyttes som udgangspunkt for oprettelsen af dokumentvektorerne. På figur 5.7 er et sådan matrice gengivet i figurens højre side.
På figur 5.7 vil dokumentvektoren $\overrightarrow{d_2}$, der alene indeholder termen „*psykose*", antage en retning identisk med aksen for denne term og en længde svarende til *én* forekomst af termen. Dokumentvektoren $\overrightarrow{d_3}$, som indeholder termerne „*psykose*" og „*skønne*", men ikke termen „*ikke*", fremtræder derfor i en retning diagonalt mellem akserne „psykose" og „skønne" og en længde svarende til én forekomst af hver af de to termer. $\overrightarrow{d_3}$ der indeholder alle tre termer er beskrevet ved retningen diagonalt på samtlige akser og længden svarende til én forekomst af hver term. Med kendskab til den pythagoræiske læresætning afslører figuren at dokumentet d_2 ligner dokumentet d_3 mere end d_1. Dette forhold underbygges af termfrekvens-matricet, hvor termen „*psykose*" indeholdt i d_2 udgør halvdelen af termerne i d_3, imens den kun udgør 1/3 af termerne i d_1.
Det valgte eksempel er stærkt simplificeret. Med et vokabularium bestående af mange termer bliver vektorrummet naturligvis ekstremt komplekst. Similaritetsbestemmelsen som beskrives i afsnit 5.4.7 er dog fortsat en triviel regneopgave takket være den geometriske repræsentation af dokumenterne.



	d_1	d_2	d_3
ikke	1	0	0
psykose	1	1	1
skønne	1	0	1

Figur 5.7: Vektorrum givet ved termfrekvens-matricets termer for dokumenterne d_1, d_2 og d_3.

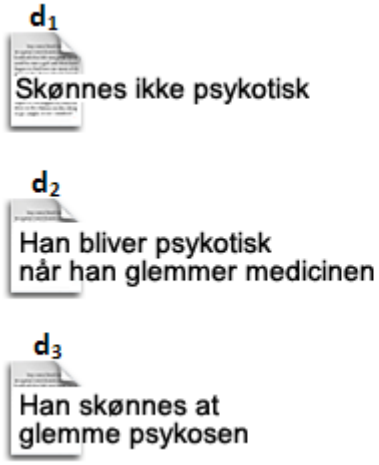
5.4.3 Dataselektion

Som udgangspunkt fastsættes selektionssnittet af journalnotater, der skal udgøre *dokumentet* for den enkelte patient og dennes behandlingsforløb. Selektionssnittet udgøres af den til forsøget fastlagte parameter der ønskes undersøgt. Eksempelvis vil undersøgelsen af parameteren „patientens sidste 10 journalnotater" føre til oprettelsen af et dokument for hver enkel patient, der hver især består af behandlingsforløbets sidste 10 journalnotater.
De dokumenter, der således fastlægges ved dataselektionen, siges at udgøre *dokumentsættet* for undersøgelsen af den valgte parameter.

5.4.4 Dataforbehandling

Efter selektion af dokumenter til data mining processen, overføres de anvendte termer i hvert dokument til et samlet vokabularium udtrykt ved et termfrekvens-matrice [Berry *et al.*, 1999]. Opbygningen af et sådant matrice er eksemplificeret på figur 5.8. På figuren er angivet dokumenterne d_1, d_2 og d_3, som hver indeholder en række termer bestemt ved dokumentvektorerne $\overrightarrow{d_1}$, $\overrightarrow{d_2}$ og $\overrightarrow{d_3}$. Termfrekvens-matricet fremkommer ved at tilføje alle dokumenternes termforekomster til matricet og derpå angive den enkelte terms forekomst for hvert dokument. For angivelse af eksempelvis dokumentet d_2 i termmatricet tages udgangspunkt i dokumentets termer, der udtrykker sætningen „*Han bliver psykotisk, når han glemmer medicinen*". Dokumentets benyttede termer er således „*han*", der optræder to gange samt termerne „*bliver*", „*psykotisk*", „*når*", „*glemmer*" og „*medicinen*", der hver optræder en gang. Termfrekvens-matricet i kolonnen for d_2 udfyldes således med angivelse af det antal gange en term forekommer. Forekomst af øvrige termer der således ikke optræder i det pågældende dokument angives med værdien 0.
Termfrekvens-matricet benyttes i den videre data mining proces til bestemmelse af

dokumenternes similaritet.



	d_1	d_2	d_3
at	0	0	1
bliver	0	1	0
glemme	0	0	1
glemmer	0	1	0
han	0	2	1
ikke	1	0	0
medicinen	0	1	0
når	0	1	0
psykosen	0	0	1
psykotisk	1	1	0
skønnes	1	0	1

Figur 5.8: Dokumenterne d_1, d_2 og d_3 er beskrevet i et termmatrice på baggrund af termforekomster i dokumenterne.

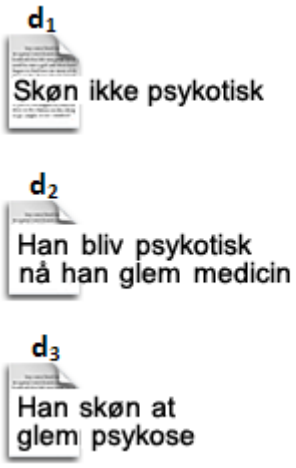
I det viste eksempel er vokabulariet givet med få termer. Inddrages flere og langt større dokumenter bliver vokabulariet hurtigt ekstremt komplekst. Der findes derfor en række metoder til at reducere vokabulariets størrelse og dermed medvirke til at reducere *støj*, hvor støj i denne sammenhæng skal ses som de termer der forekommer på tværs af samtlige dokumenter og som derfor ikke kan benyttes til at skelne mellem de enkelte dokumenter. Sådanne metoder inkluderer *lemmatisering*, *fjernelse af stop-ord* og *uddragning af domænespecifikke termer*.

Lemmatisering

Lemmatisering henviser til uddragning af en fælles form for den enkelte term [Carlberger *et al.*,]. En sådan form kan være grundstammen ved bøjninger af verber, som for eksempel termen „*gå*“, der kan optræde i dokumenterne som „*gå*”, „*går*” og „*gik*”. Med udgangspunkt i eksemplet på figur 5.8 vil uddragning af grundstammen for de enkelte termer føre til det viste vokabularium på figur 5.9. Lemmatiseringen har i eksemplet den effekt, at de to felter med de oprindeligt forskellige forekomster af grundstammen „*glem*” for termerne „*glemme*” og „*glemmer*” fra figur 5.8 nu kan aggregeres.

Fjernelse af stop-ord

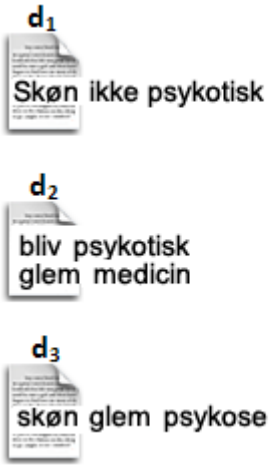
Fjernelse af stop-ord henviser til en reduktion af vokabulariet, hvor ord uden betydning; de såkaldte *stop-ord* forsøges fjernet helt. Højfrekvente termer, det vil sige termer, der forekommer på tværs af alle dokumenter, vil således udgøre dokumentsættets *stopordsliste* [Berry *et al.*, 1999]. Eksempler på ord fra en sådan liste er ord som „at”, „er” og „som”, der i sig selv ikke tilfører egentlig betydning til et dokument,



	d_1	d_2	d_3
at	0	0	1
bliv	0	1	0
glem	0	1	1
han	0	2	1
ikke	1	0	0
medicin	0	1	0
nå	0	1	0
psykose	0	0	1
psykotisk	1	1	0
skøn	1	0	1

Figur 5.9: Lemmatisering af vokabulariet hvormed kun termernes grundformer uddrages.

idet de indgår i stort set alle dansksprogede dokumenter. Beregning af similaritet mellem dokumenter, hvor sådanne ord indgår, kan således siges i højere grad at være påvirket af støj. Fjernelse af stop-ord tjener dermed til støjreduktion og er illustreret på figur 5.10, hvor stop-ordene „*at*", „*han*" og „*nå*" fra figur 5.9 er blevet fjernet.



	d_1	d_2	d_3
bliv	0	1	0
glem	0	1	1
ikke	1	0	0
medicin	0	1	0
psykose	0	0	1
psykotisk	1	1	0
skøn	1	0	1

Figur 5.10: Termfrekvens-matricet efter fjernelse af stop-ordene „*at*", „*han*" og „*nå*" fra figur 5.9.

Uddragning af domænespecifikke termer

Uddragning af domænespecifikke termer henviser til eksklusion af samtlige termer fra termfrekvens-matricet hvor fællesformen er identificerbar. Uddragning af domænespecifikke termer kan således betragtes som en slags *invers lemmatisering*, der fører til et vokabularium alene bestående af uidentificerbare termer. Eksempelvis vil den

danske retskrivningsordbog kunne anvendes på ord som „*sygdom*" eller „*mani*", mens langt mere specifikke kliniske termer som „*dystoni*" og „*dyslipidæmi*"ikke vil være identificerbare.
Argumentet for anvendelsen af denne metode er, at den sikrer inklusion af domænespecifikke ord, som ikke skønnes identificerbare ved den anvendte lemmatiseringsmetode. Argumentet imod metodens anvendelse er imidlertid, at den gør det valgte datasæt ekstremt følsom overfor støj eksempelvis i form af stavefejl.

5.4.5 Datatransformation

Datatransformation indenfor tekst data mining refererer til transformationen af vokabulariet udtrykt ved dokumentvektorerne samt til bestemmelsen af similariteten mellem dokumentvektorerne.
Som nævnt danner vokabulariet grundlag for udtrykket af en vektor ved bestemmelse af en såkaldt *termvægt* for den enkelte term $t_{i,j}$. Termvægten udtrykker betydningen af den enkelte term for det pågældende dokument d_i og det er således dokumentets vægtede termer $w_{i,j}$, der udtrykker vektoren som angivet i 5.2 [Buckley , 1993].

$$\overrightarrow{d_i} = (w_{i,1}, w_{i,2}, ..., w_{i,n}) \tag{5.2}$$

Metoder til termvægtning vil blive gennemgået i afsnit 5.4.6 mens similaritetsbestemmelse på baggrund af termvægtningen vil blive gennemgået i afsnit 5.4.7.

5.4.6 Termvægtning

I eksemplet på figur 5.7 i det 3-dimensionelle rum var de valgte termer *ens* vægtede, hvilket vil sige, at ingen af vokabulariets termer blev anset for at være af større eller mindre betydning end de øvrige. Termvægtning anvendes til at indføre en skelnen mellem termernes betydning i form af de såkaldte termvægte, som er en nummerisk værdi, der knyttes til hver enkelt term. Formålet med termvægtningen er således at influere på dokumentvektorerne, hvor det i højere grad er de højt vægtede termer, der er udslagsgivende for dokumentvektorernes placering i vektorrummet end det er de lavt vægtede termer.
I det følgende præsenteres en række metoder til at bestemme en vægt til datasættets termer:

Maksimal termfrekvens (maxtf)

Maksimal termfrekvens (maxtf) vægtning udregnes for den enkelte term $t_{i,j}$ for dokumentet d_i ved at dividere antallet af forekomster af den enkelte term (termfrekvensen $tf_{i,j}$) med dokumentets højeste termfrekvens $maxtf_i$, som vist i 5.3 [Martínez-Fernández *et al.*, 2004].

$$w_{i,j} = \frac{tf_{i,j}}{maxtf_i} \tag{5.3}$$

Der er således tale om en normaliseret vægt, idet den højeste termfrekvens for hvert dokument tages som udgangspunkt i beregningen, hvorved længere dokumenter med deraf følgende flere termer ikke alene på grund af det forholdsmæssigt større antal ord opnår højere værdier end dokumenter med færre termer.

Termfrekvens-invers dokumentfrekvens (tf-idf)

Termfrekvens-invers dokumentfrekvens (tf-idf) vægtning er et statistisk mål, der tager udgangspunkt i *betydningen* af den enkelte term for dokumentet. Termens betydning fastlægges ved at tage logaritmen til divisionen mellem antallet af dokumenter N og det antal dokumenter n, hvori den enkelte term $t_{i,j}$ optræder som vist i 5.4 [Liu and Loh, 2007].

$$w_{i,j} = tf_{i,j} * log(\frac{N}{n}) \tag{5.4}$$

Der er således tale om en favorisering af de termer der optræder i få dokumenter, imens termer, der forekommer i flere dokumenter, tildeles en lavere vægt.

Term *boosting*.

Term *boosting* henviser til tildelingen af en kunstig og væsentlig højere termvægt til en given term. Dette har til formål at øge en given terms betydning for similariteten mellem dokumenterne i datasættet.

5.4.7 Similaritetsmål

Efter valg af termvægtning benyttes similaritetsmålet til at fastlægge similariteten mellem samtlige dokumenter i datasættet. Similariteten givet ved en nummerisk værdi kan efterfølgende benyttes af clustering-metoden til opdeling af dokumenterne i clusters med størst lighed.
Flere similaritetsmål er beskrevet i litteraturen [Kowalski, 1997], men jeg har dog udvalgt to hyppigt forekommende mål; *cosinus* og *Jaccards similaritetskoefficient*, som begge har den egenskab, at de er normaliserede og således resulterer i en værdi i intervallet [0,1], hvor 0 er det absolutte fravær af lighed og 1 beskriver den absolutte lighed. Værdier mellem 0 og 1 angiver den procentuelle lighed mellem de to dokumenter.

Cosinus similaritetsbestemmelse

Cosinus similaritetsbestemmelse benyttes til udregning af cosinus til vinklen mellem to dokumentvektorer. Formlen for udregning af cosinus mellem $\overrightarrow{d_1}$ og $\overrightarrow{d_2}$ er vist i 5.5 [Berry *et al.*, 1999].

$$cos(\overrightarrow{d_1}, \overrightarrow{d_2}) = \frac{(\overrightarrow{d_1}, \overrightarrow{d_2})}{\parallel \overrightarrow{d_1} \parallel * \parallel \overrightarrow{d_2} \parallel} = \frac{\sum_{i=1}^{n} d_1 d_2}{\sqrt{\sum_{i=1}^{n} d_1^2}\sqrt{\sum_{i=1}^{n} d_2^2}} \tag{5.5}$$

Jaccards similaritetskoefficient

Jaccards similaritetskoefficient er an anden model for similaritetsbestemmelsen mellem dokumenter. Koefficienten udregnes ved at dividere antallet af termer, de to dokumenter har til fælles, med det samlede antal termer i dokumenterne, der sammenlignes. Formlen for udregning af Jaccards similaritetskoefficient mellem vokabularierne d_1 og d_2 er vist i 5.6 [Romesburg, 2004].

$$Jaccard(d_1, d_2) = \frac{\mid d_1 \cap d_2 \mid}{\mid d_1 \cup d_2 \mid} \tag{5.6}$$

5.5 Afrunding

I dette kapitel er data mining processen blevet gennemgået med særligt fokus på tekst data mining, der anvendes ved uddragning af viden fra ustruktureret tekst som tilfældet er med hovedparten af de patientinformationer i form af journaldata, der registreres i EPJ-systemet.
Det følgende kapitel indeholder specialets metodebeskrivelse med særlig fokus på data mining processen og dens anvendelse i specialets undersøgelser.

Kapitel 6

Metode

Specialets forsøg har til formål at efterprøve hypotesen, at der i løbet af en given behandlingsperiode registreres informationer i EPJ, der kan bruges som parametre for udfaldet af den pågældende behandling.

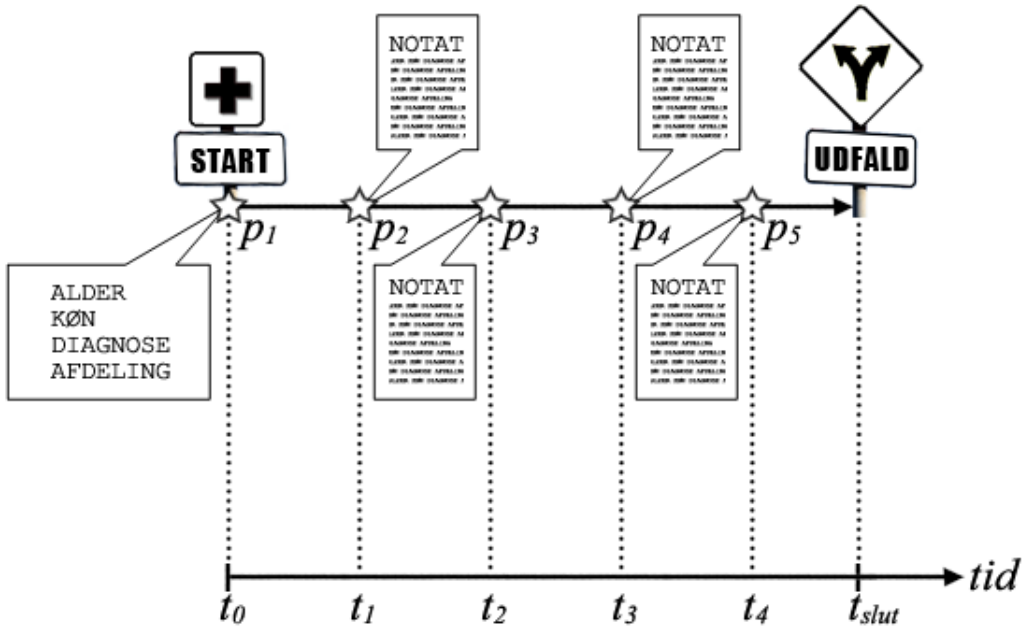
For at bestemme betydningen af en eller flere parametre i EPJ-systemet, har jeg valgt at anvende clusteranalyse som data mining-metode på et datasæt bestående af de valgte parametre, som de optræder i de til forsøget udvalgte behandlingsforløb. Behandlingsforløbene er udvalgt på baggrund af udfaldet af de enkelte behandlingsforløb. Disse udfald benævnes forsøgets *udfaldsparametre*. Udfaldsparametrene udgøres af to klinisk set diametralt modsatte behandlingsudfald, der er gensidigt ekskluderende: *Behandlingsskift* og *udskrivning*.

De valgte parametre udgøres af henholdsvis de journalførte oplysninger i EPJ-systemet fra behandlingsforløbets start til dets ophør samt af de kliniske variable, der er tilstedeværende i udgangspunktet for det af klinikeren valgte behandlingsforløb. Ovennævnte forhold er illustreret på figur 6.1.

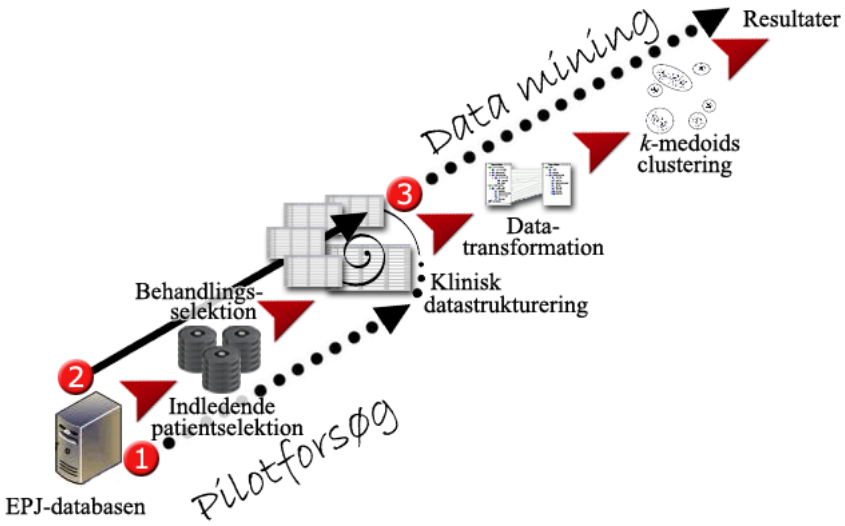
Reliabiliteten af clusterinddelingen er fastlagt som Kappa koefficienten mellem *udfaldet* af clusteranalysen og gruppeinddelingen baseret på de på forhånd valgte *udfaldsparametre*.

6.1 Metodeopbygning

Forsøgsopbygningen i specialet stemmer i store træk overens med den generelle data mining procesmodel fremstillet i afsnit 5.2. På figur 6.2 er forsøgsopbygningen, som den tager sig ud i specialet illustreret. Metodeafsnittet er opbygget som en beskrivelse af samtlige trin i forsøgsopbygningen, der overordnet set er opdelt i (1) et pilotforsøg, (2) parameterselektion ved de enkelte data mining processer og (3) de enkelte data mining-processer, som igen er baseret på tre typer: et tekst data mining forløb, et data mining forløb baseret på kliniske variable samt et data mining forløb baseret på koblingen mellem de to foregående.



Figur 6.1: Illustration af de journalførte oplysninger i EPJ-systemet, som de optræder i forbindelse med et behandlingsforløb fra forløbets start (t_0) til dets ophør (t_{slut}). Både journalnotater (repræsenteret ved parametrene p_2-p_5) og kliniske variable (p_1) forventes at dække over markører for det pågældende behandlingsudfald.



Figur 6.2: Forsøgsopbygningen i specialet; (1) betegner den indledende patientselektion og udgøres af pilotforsøget; (2) betegner dataselektionen på baggrund af den kliniske datastrukturering afstedkommet af pilotforsøget og (3) betegner den egentlige data mining proces med undersøgelse af udvalgte parametres betydning for de valgte behandlingsudfald.

6.2 Forsøgsdesign

6.2.1 Valg af inklusionskriterier

Jeg har i specialets afsnit 2.5 opstillet en række inklusionskriterier for selektion af EPJ-registrerede patientdata til specialets forsøg. De valgte inklusionskriterier er gengivet herunder:

- Patientgrundlaget udgøres af patienter med afsluttede indlæggelser registreret på sengeafsnit på Psykiatrisk Center Sct. Hans i perioden fra 1. april 2003 til 1. april 2008
- Patienter skal have minimum en af følgende diagnoser: F1, F2 eller F3
- Patienter bidrager med journalførte oplysninger fra behandlingsforløb registreret under patientens *første* indlæggelse
- Behandlingsforløb skal udgøres af minimum 4 ugers antipsykotisk monoterapi
- Behandlingsforløb skal føre til enten behandlingsskifte eller udskrivning; for behandlingsforløb førende til behandlingsskifte skal der være tale om indlæggelsens første behandlingsforløb
- Patienten bidrager med kun *et* behandlingsforløb. Hvis patientens indlæggelse kan bidrage med flere behandlingsforløb til undersøgelsen selekteres alene det første behandlingsforløb

At der skal være tale om indlæggelse på sengeafsnit skyldes, at alene sådanne indlæggelser dokumenteres dagligt i EPJ-systemet. Indlæggelsen skal endvidere være afsluttet, fordi al potentiel brugbar viden om det samlede indlæggelsesforløb skal være til stede ved forsøgets start.
Valget af periodeafgrænsningen skyldes hensynet til at kunne indsamle så komplette informationer for så lang en periode som muligt.
Diagnosekravet skal være opfyldt for at sikre, at der er tale om antipsykotisk behandling.
For at mindske bias-faktorer fra både patienter og behandleres side, har jeg valgt at undersøgelsen kun skal inddrage første indlæggelse i EPJ-systemet. Ligeledes skal der ved selektion af behandlingsforløb frem til udfaldet behandlingsskifte være tale om første behandlingsforløb.
Valget af monoterapi er truffet for at sikre så bred teoretisk baggrundsviden for forsøgsudfald som muligt. At behandlingen skal være foregået i minimum 4 uger er valgt for at sikre, at medicineringen for det pågældende behandlingsforløb har haft indflydelse på behandlingsudfaldet.

6.3 Pilotforsøg

At jeg har valgt at tage udgangspunkt i et pilotforsøg, hænger sammen med at det stort set ikke har været muligt at opnå kendskab til det tilgængelige patientmateriale før specialets påbegyndelse. Den anden væsentlige pointe ved pilotforsøget er at datastrukturen i EPJ-systemet fordrer en indledende transformation, der tillader mig at bestemme de udfaldsparametre som både identifikation og data mining af variable i EPJ-systemet er baseret på.
Pilotforsøget har således til formål at bestemme de patienter registreret i EPJ-systemet, der vil fungere som forsøgets patientkohorte samt at konstruere tabeller på baggrund af den eksisterende struktur i EPJ-systemets database til brug for den videre data mining proces.
Til pilotforsøget har jeg udviklet en applikation til dataindhentning fra EPJ-systemet. I første omgang anvendes applikationen til den indledende patientselektion. Ved alle identifikationer, der udføres af applikationen, har jeg valgt at oprette tabeller i en database uden personhenførbare oplysninger, der således giver mulighed for hurtigere tilgang til de relevante oplysninger i specialets forsøg samtidig med at data anonymiseres.

6.3.1 Indledende patientselektion

Da jeg vurderer, at antipsykotisk monoterapi er størst betydende faktor i forhold til selektionen af patientmaterialet, har jeg for pilotforsøget valgt i første omgang at identificere patienter med forekomst af monoterapi i minimum 4 uger indenfor perioden fra 1. april 2003 til 1. april 2008.

Antipsykotika

I specialet har jeg anvendt [Pedersen *et al.*, 2008]'s antipsykotikafortegnelse. Disse antipsykotika er gengivet i tabel 3.5 og tabel 3.6. På baggrund af denne liste har jeg oprettet tabellen DRUG, DRUG_VAR samt DRUG_GEN. Indeholdet af disse tre tabeller er beskrevet i tabel 6.1.
Tabelkonstruktionerne har indledningsvis til formål at identificere samtlige antipsykotika-ordinationer i EPJ-systemet samt at gruppere forskellige stavemåder for samme præparat for entydigt at kunne bestemme sådanne forekomster. Selvom præparatnavnet optræder under handelsnavnet, har jeg endvidere valgt at oprette en tabel for præparaternes generiske navne. For samtlige præparatorienterede tabeloprettelser gælder, at der alene er tale om antipsykotika. Figur 6.3 angiver E/R-diagram for tabeloprettelserne.

Monoterapi

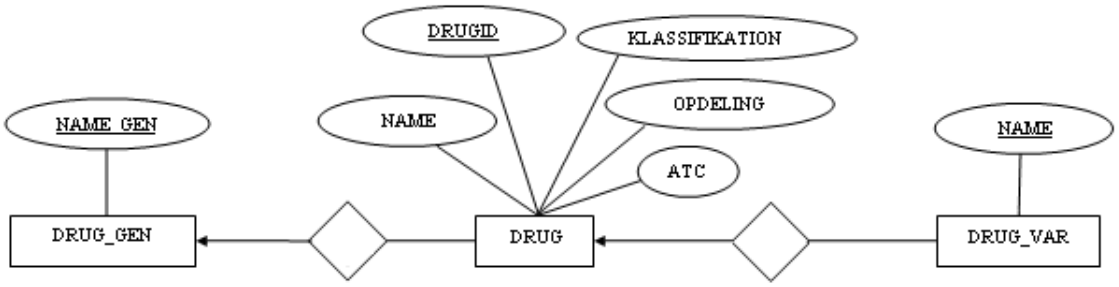
Monoterapi er tidligere defineret som den behandling, hvor patienten modtager netop ét antipsykotikum. Idet der som minimum skal være tale om 4 ugers sammenhængende monoterapi indenfor den valgte periode, har jeg oprettet en applikation, der på

DRUG	
DRUGID	Løbenummer for præparat
NAME	Præparatnavn (handelsnavn)
KLASSIFIKATION	Typisk eller atypisk antipsykotikum
ATC	ATC-koden for præparatet

DRUG_VAR	
NAME	Stavemåde for præparatet i EPJ-systemet

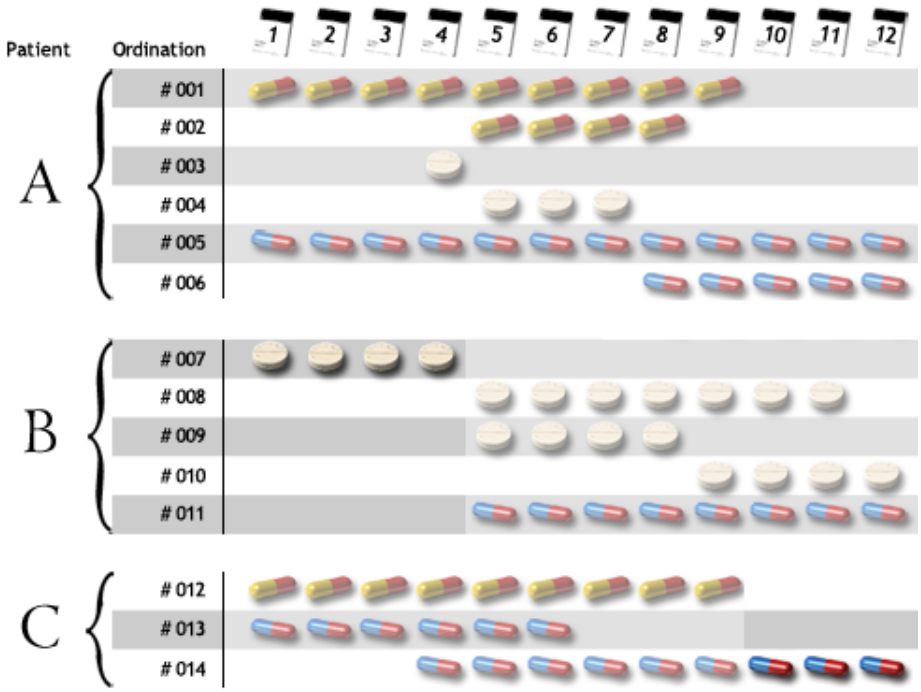
DRUG_GEN	
NAME_GEN	Præparatets generiske navn

Tabel 6.1: Tabeller oprettet til at håndtere præparatforekomster i EPJ-systemet, aflede korrekt stavemåde samt gruppere præparater med samme aktivstof på baggrund af præparaternes generiske navne.



Figur 6.3: E/R-diagram for de afledte præparattabeller DRUG, DRUG_VAR og DRUG_GEN.

baggrund af forespørgsler til EPJ-databasen gennemløber hver enkelt dato og sammenregner varigheden af monoterapiforløb for hver patient. Figur 6.4 illustrerer 12 dages ordinationer for patienterne A, B og C. Ved gennemløb af dette eksempel vil patient A figurere med 0 dage i monoterapi, patient B med 4 dage (fra 1. til 4. dagen af ordination #007) og patient C med 3 dage (fra 10. til 12. dagen af ordination #014).



Figur 6.4: Skitseret forløbsgennemgang af ordinationer for patienterne A, B og C i en periode af 12 dage (angivet med kalenderblade øverst på figuren). Ordinationerne for patient A er nummereret fra #001 til #006, for patient B er det ordinationerne #007 til #011 og for patient C ordinationerne #012 til #014. Denne nummerering har til hensigt at eksemplificere måden, hvorpå præparatordinationer registreres i databasen i EPJ-systemet, hvor alle ordinationer registreres med et unikt løbenummer. I dette eksempel udgør hver specifik nummerering én unik ordination. Ved vurdering af monoterapi-forløb for patienterne A, B og C vil det kun være patient B med ordinationen #007 i dagene 1 til 4 og patient C med ordinationen #014 i dagene 10 til 12, hvor der er tilfælde af monoterapi.

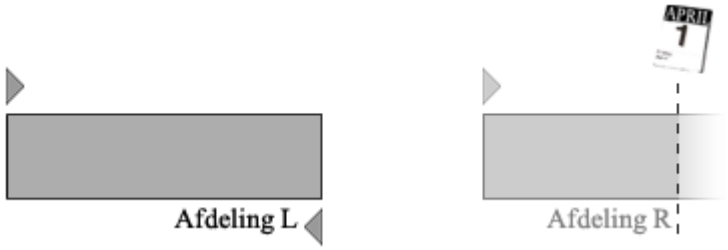
Indlæggelsesforløb

Indlæggelser registreres på afdelingsniveau, hvilket betyder at der ved identifikation af den samlede indlæggelse for en given patient skal tages højde for den tidslige adskillelse mellem flere på hinanden følgende afdelingsforløb. En sådan tidslig adskillelse kan opstå enten ved overflytning mellem afdelinger eller ved temporær udskrivelse i forbindelse med korte ophold udenfor centret i anden varetægt. Da jeg ønsker, at begge sådanne tilfælde skal fastholde patienten som indlagt på centret, er det nødvendigt at tillade en midlertidig pause mellem to på hinanden følgende afdelingsforløb. Da

det er ofte forekommende, at patienter eksempelvis får lov at tage hjem og holde jul med familien, har jeg valgt at tillade et ophold mellem afdelingsindlæggelserne på maksimalt 9 dage.
På samme måde kan der ved overflytning mellem to afdelinger optræde et tidsligt overlap. På figur 6.5 er et samlet indlæggelsesforløb forsøgt illustreret med et temporært ophold mellem første og anden afdelingsindlæggelse samt et overlap mellem de to afsluttende afdelingsforløb. Eventuelle afdelingsforløb placeret mere end 9 dage på begge sider af de fremhævede afdelingsforløb medregnes ikke i indlæggelsen.

Figur 6.5: Et samlet indlæggelsesforløb (fremhævet) bestående af tre afdelingsforløb startende på afdeling L efterfulgt af et forløb på afdeling M og slutteligt på afdeling R. Afdelingsforløbene på afdeling R og afdeling M er tidsligt adskilt, men grupperes dog fortsat som samme indlæggelsesforløb.

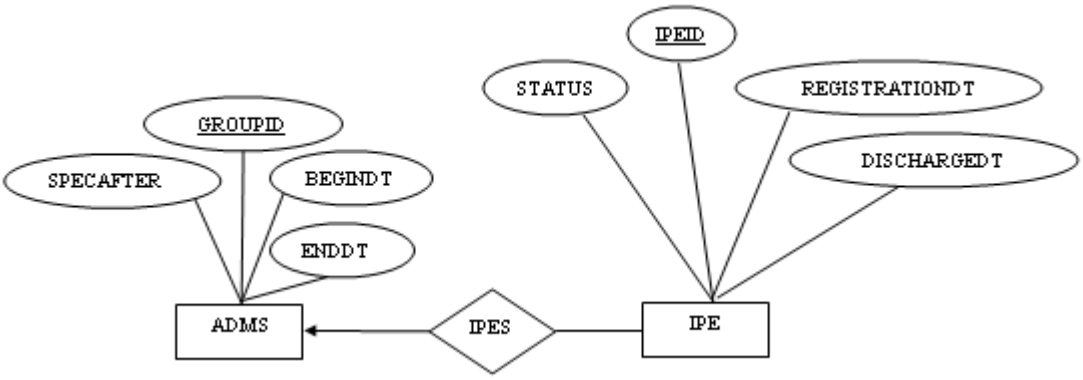
Af figur 6.5 ses, at de tre centrerede afdelingsforløb er blevet forbundet i et samlet indlæggelsesforløb. Indlæggelsens indskrivnings- og udskrivningsdato er symboliseret ved de store pile, imens de små pile blot angiver de indskrivnings- og udskrivningsdatoer for de afdelingsforløb, der indgår i indlæggelsen. Et forudgående og et efterfølgende afdelingsforløb er skitseret på hver side af indlæggelsesforløbet for at illustrere, at afstanden gør, at disse ikke indberegnes i den fremhævede indlæggelse. For alle tre afdelingsforløb gælder, at indskrivnings- og udskrivningsdatoer ligger inden for en afstand af 9 dage. På figuren ses endvidere, at afdelingsforløbet for afdeling M og afdeling R overlapper hinanden. Dette illustrerer det forhold, at registrering af udskrivnings- og indskrivningsdatoer varetages af de respektive afdelinger, hvorfor der kan forekomme sådanne dato-overlap. Bemærk at et indlæggelsesforløb ikke skal indeholde flere på hinanden følgende afdelingsforløb for at få betegnelsen indlæggelsesforløb. Således kan de to isolerede afdelingsforløb (på grund af en tidslig adskillelse på mere end 9 dage) på figur 6.6 betragtes som to indlæggelsesforløb: et afsluttet på afdeling L og et igangværende på afdeling R.
Specialambulatoriet skiller sig ud fra de øvrige afdelinger ved ikke at indeholde nogen sengeafsnit. Jeg har derfor valgt, at indskrivning på Specialambulatoriet ikke medregnes som en indlæggelse. Dette betyder, at jeg har valgt at lade applikationen registrere indlæggelsesforløb, der overgår til indskrivning på Specialambulatoriet, som ophørt. Dette er eksemplificeret i figur 6.7.
Til registrering af de identificerede indlæggelsesforløb opretter jeg tabellerne ADMS og IPES. Tabellen ADMS benyttes om entiteten indlæggelse med attributterne indlæggelsesdato og udskrivningsdato, imens tabellen IPES benyttes om relationen mellem indlæggelsen og den enkelte afdelingsindlæggelse (IPE) som beskrevet i tabel



Figur 6.6: Et afsluttet afdelingsforløb på afdeling L (fremhævet) efterfulgt af et på skæringstidspunktet igangværende afdelingsforløb på afdeling R.

Figur 6.7: Et afdelingsforløb på afdeling R efterfølges af et afdelingsforløb på specialambulatoriet. Da specialambulatoriet ikke klassificeres som en indlæggelse, er det alene afdelingsforløbet på afdeling R, der indgår i indlæggelsesforløbet (fremhævet).

6.2. Figur 6.8 angiver E/R-diagram for tabeloprettelserne.



Figur 6.8: E/R-diagram for afledte indlæggelsestabeller ADMS og IPES, hvor sidstnævnte fungerer som koblingen mellem den samlede indlæggelse (ADMS) og afdelingsindlæggelserne (IPE) i den eksisterende database struktur for EPJ-systemet.

6.3.2 Klinisk datastrukturering

Som supplement til patientselektionen har jeg valgt at udvide min applikation med et identifikationsmodul til bestemmelse af *behandlingsforløb* og *diagnoser*.

ADMS	
GROUPID	Løbenummer for indlæggelsen
BEGINDT	Indlæggelsesdato
ENDDT	Udskrivningsdato
SPECAFTER	Angivelse af om indlæggelsen afsluttes med indskrivning på specialambulatoriet

IPES	
GROUPID	Løbenummer for indlæggelsen
IPEID	Løbenummer for afdelingsindlæggelsen

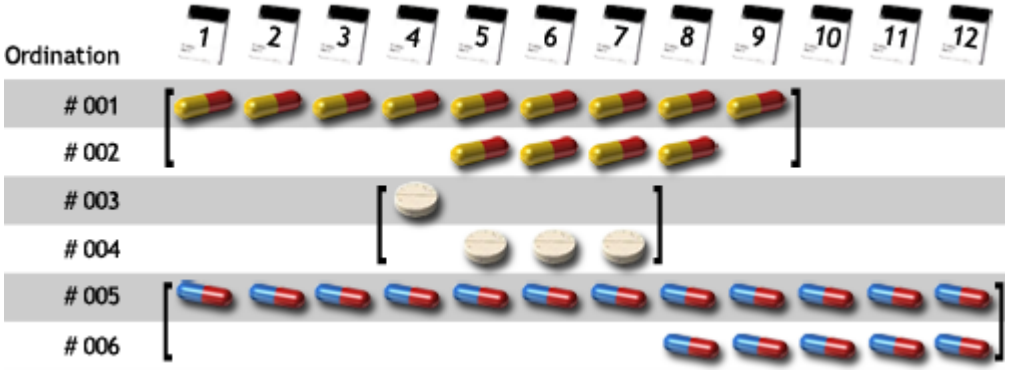
Tabel 6.2: Tabeller oprettet med henblik på at gruppere afdelingsindlæggelser i et samlet indlæggelsesforløb (ADMS).

Behandlingsforløb

For at skelne mellem forskellige perioder i en patients ordinationsforløb, har jeg valgt at benytte betegnelsen *behandlingsforløb*, der er en væsentlig abstraktion i forhold til den form, efter hvilken præparatordinationer registreres i EPJ-systemet. Før det er muligt at klassificere forekomster af forskellige behandlingsforløb, er jeg imidlertid nødt til at addressere et væsentligt problem ved registrering af præparatordinationer i EPJ-systemet nemlig fejlraten.

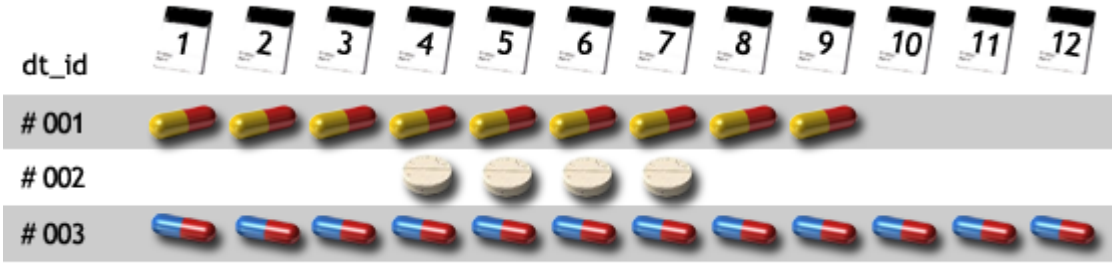
Fejlraten hvormed informationerne er blevet registreret i EPJ-systemet forekommer forbavsende høj. Eksempelvis optræder seponeringsdatoen for en ordination ikke sjældent før ordinationens begyndelsdato, ligesom der til til tider optræder flere samtidige ordinationer af et og samme præparat i EPJ-systemet. Dette ses blandt andet ved ændring af dosisstørrelse for en given ordination, hvor den øgede dosisstørrelse fejlagtigt registreres som en selvstændig ordination i stedet for en ændring af den eksisterende ordination. Før en opdeling i behandlingsforløb kan finde sted, er det derfor nødvendigt indledningsvist at sammenkæde sådanne sammenfaldende ordinationer samtidig med, at der tages højde for datomæssige fejl. Idet dosisstørrelser først er registreret fast fra midten af 2005 i EPJ-systemet, har jeg valgt ikke at medtage dosisstørrelser, men alene at betragte datoen for ordinationens forekomst.

En grafisk fremstilling af problemstillingen er eksemplificeret på figur 6.9, der tager udgangspunkt i ordinationskemaet for patient A på figur 6.4. På figur 6.9 er klynger af samme ordination fremhævet. Sådanne klynger skyldes flere samtidige ordinationer af samme præparat, hvilket er hyppigt forekommende i EPJ-systemets registreringer, hvor brugere nogle gange ved dosisjustering i stedet for ændring af den eksisterende ordination eksempelvis *opjusterer* dosis ved at tilføje samme præparat. Ligeledes observeres enkelte samtidige forekomster af både konventionel medicinering og depotmedicinering med samme præparat. I alle tilfælde gælder, at der *ikke* er tale om polyfarmaci, hvis flere samtidige ordinationer alene udgøres af præparater



Figur 6.9: Med udgangspunkt i behandlingsforløbet for patient A på figur 6.4 illustreres en grundigere gennemgang af patientens præparatordinationer. Figuren eksemplificerer en gennemgang af præparatordinationerne, hvor det konstateres, at ordinationerne #001 og #002 omfatter *identiske* præparater, ligesom ordinationerne #003 og #004 samt ordinationerne #005 og #006. Disse ordinationer er derfor grupperet i forløbets længderetning (omgivet af kantede parenteser).

med samme generiske navn[1]. Jeg har derfor valgt at lade min applikation *gruppere* sådanne forekomster i tabellen DRUGTREATMENTSETS, som jeg har oprettet til dette formål. På figur 6.10 er præparatordinationerne således aggregeret, hvor hver afledt ordination er tildelt et særligt løbenummer (dt_id).



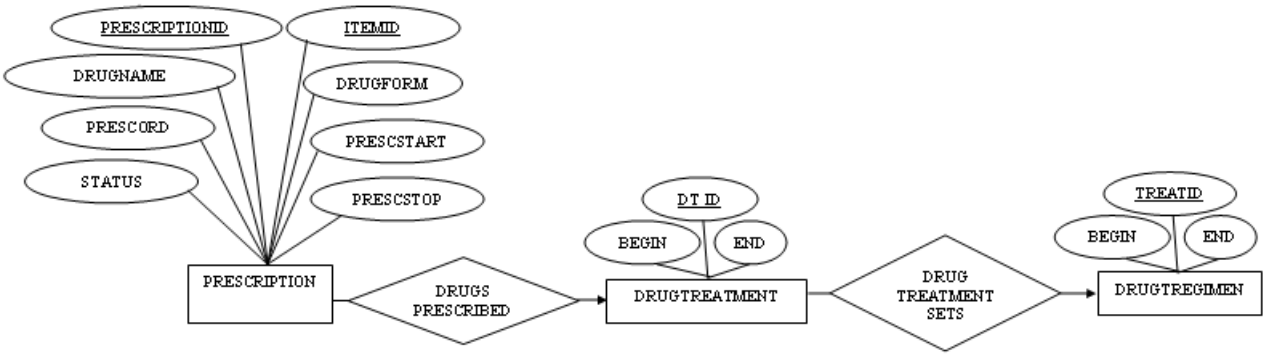
Figur 6.10: Med udgangspunkt i behandlingsforløbet for patient A på figur 6.9 med de grupperede præparatordinationer, er disse ordinationer nu aggregeret til blot at omfatte tre ordinationer mod tidligere seks forskellige ordinationer. I stedet for et ordinationsnummer, er ordinationerne tildelt et behandlingsløbenummer (dt_id).

Figur 6.11 illustrerer, hvordan de grupperede ordinationer (angivet med løbenummeret dt_id) fra figur 6.10 nu kan grupperes i specifikke *behandlingsforløb*. Hvert behandlingsforløb er på figuren omgivet af en øvre og en nedre kantet parentes, der udtrykker de kontinuum af samme medicinske behandling.
Jeg har udvidet min applikation med et modul, der registrerer forekomster af behandlingsforløb (angivet med løbenummeret treat_id) og tilføjer dem til tabellen DRUGTREATMENTSETS. Dette leder til oprettelsen af tabellerne angivet i tabel 6.3,

[1]Se afsnit 3.3.1.

Figur 6.11: De aggregerede ordinationsforløb (dt_id) på figur 6.10 for patient A er her opdelt i forskellige perioder af kontinuum, hvor ophør eller tilføjelse af yderligere præparater til forløbet afgrænser de forskellige perioder (omgivet af en øvre og en nedre kantet parentes). Disse perioder betegnes behandlingsforløb og er yderligere nummereret.

hvoraf E/R-diagrammet herfor er vist på figur 6.12.



Figur 6.12: E/R-diagram for de afledte behandlingsforløbstabeller DRUGTREGIMEN, DRUGTREATMENTSETS, DRUGTREATMENT og DRUGSPRESCRIBED. Sidstnævnte tabel kobler de afledte tabeller sammen med tabellen PRESCRIPTION i den eksisterende struktur for EPJ-systemets database.

Diagnoser

Fastlæggelse af diagnoser sker på baggrund af tabellerne DIAGPROB og NOTEHEADER[2]. Hvor diagnoserne i NOTEHEADER alene er angivet med diagnosens navn, er diagnoserne i tabellen DIAGPROB angivet med både navn og diagnosekode. For kategorisering af diagnosen som enten hoved- eller bi-diagnose, er registreringen foretaget i tabellen NOTES med attributten mark, dog kun gældende for diagnoser registreret i DIAGPROB.

[2]Se afsnit 4.3.1

DRUGTREGIMEN	
TREATID	Løbenummer for behandlingsforløbet
BEGIN	Startdato for behandlingsforløbet
END	Slutdato for behandlingsforløbet

DRUGTREATMENTSETS	
TREATID	Løbenummer for behandlingsforløbet
DT_ID	Løbenummer for grupperede præparatordinationer

DRUGTREATMENT	
DT_ID	Løbenummer for grupperede præparatordinationer
BEGIN	Startdato for de grupperede præparatordinationer
END	Slutdato for de grupperede præparatordinationer

DRUGSPRESCRIBED	
DT_ID	Løbenummer for grupperede præparatordinationer
PRESCRIPTIONID	Løbenummer for ordinationen

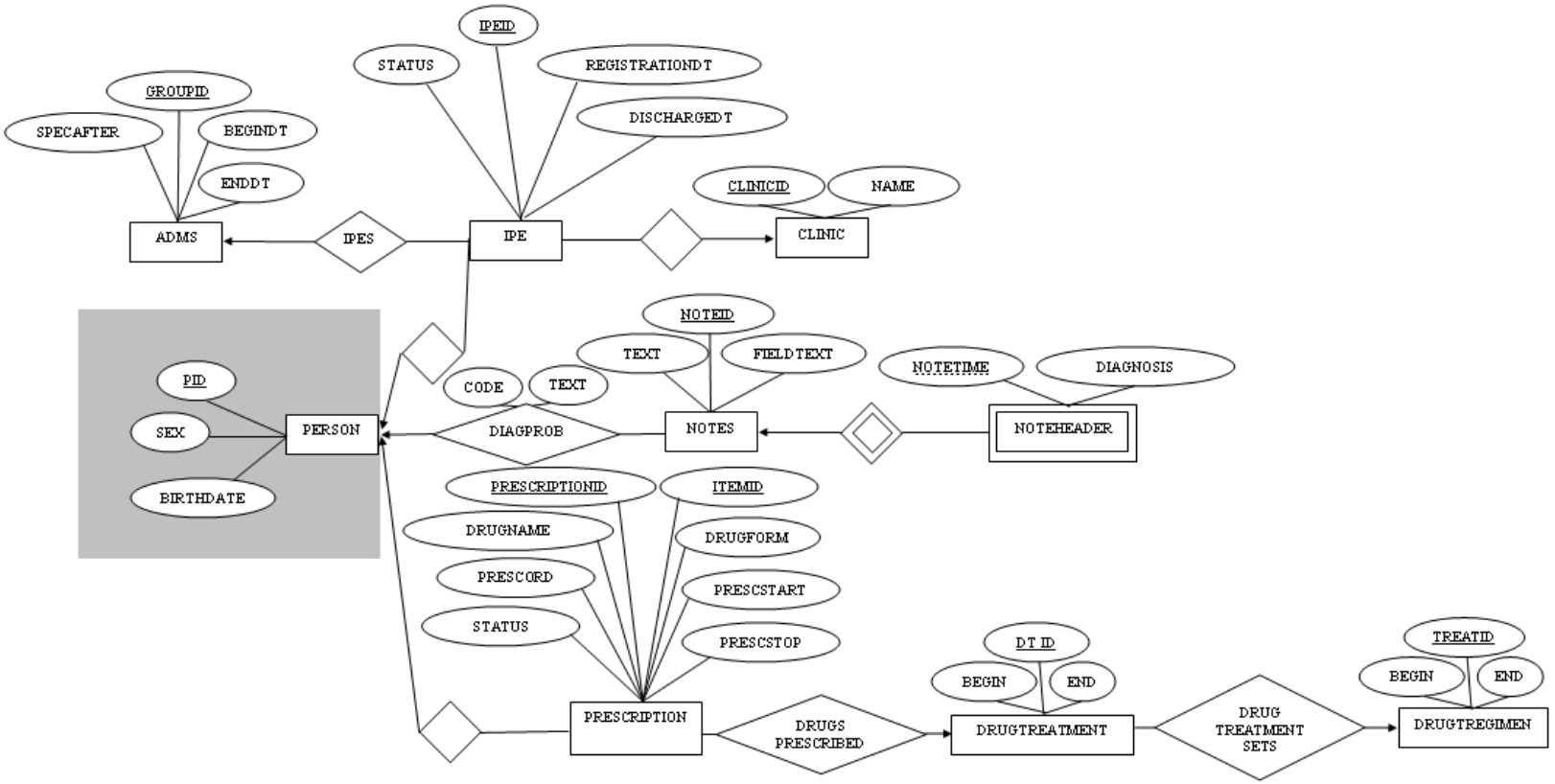
Tabel 6.3: Tabeller oprettet med henblik på at gruppere identiske ordinationer og samle alle ordinationsforekomster i afledte behandlingsforløb.

Som udgangspunkt forsøges diagnosen fastlagt på baggrund af tabellen DIAGPROB. Diagnoser, der ikke kan fastsættes på den baggrund, bestemmes ud fra tabellen NOTEHEADER, og jeg opretter den afledte tabel DIAGNOSER_UKATEGORISERET, hvori diagnosekoden registreres på baggrund af navngivningen i noteheader. Jeg har desuden valgt at oprette tabellen DIAGNOSER_KATEGORISERET med kategoriserede diagnoser for at lette adgangen til disse data i forbindelse med undersøgelse af diagnosens betydning for udfaldet. Disse tabeller er fremstillet i tabel 6.4.

6.3.3 Resultat af pilotforsøget

På baggrund af tabeloprettelserne i pilotforsøget fremkommer en samlet tabelstruktur for EPJ-systemets database som angivet på figur 6.13.
Ud af en samlet patientpopulation på 1987 patienter[3] for perioden fra 1. april 2003 til 1. april 2008 identificeres efter tabeloprettelserne ved den indledende patientselektion i afsnit 6.3.1 ialt 697 patienter med 4 ugers antipsykotisk monoterapi [4]. Der er således tale om en reduktion af periodens tilgængelige patientmateriale på knap 65%.
Tabeloprettelserne i afsnit 6.3.2 er beskrevet som den kliniske datastrukturering og fører til en klar specificering af samtlige behandlingsforløb for de identificerede 697 patienter. Fordelingen blandt F 1-3 diagnoserne for de identificerede patienter fremgår af tabel 6.5.

[3] Se bilag A.1.3
[4] Se bilag A.2.1



Figur 6.13: E/R-diagram for udsnittet af databasen med tilføjelse af de afledte tabeller, der anvendes til specialets undersøgelser. Den grå markering omkring tabellen PERSON indikerer, at der er tale om udsnittets mest centrale tabel.

DIAGNOSER_UKATEGORISERET	
TREATID	Løbenummer for behandlingsforløbet
F_DIAGNOSE	ICD-10 kode for patientens diagnose

DIAGNOSER_KATEGORISERET	
TREATID	Løbenummer for behandlingsforløbet
F_DIAGNOSE	ICD-10 kode for patientens diagnose
HOVED_BI	Markering af enten hoved- eller bi-diagnose

Tabel 6.4: Tabeller oprettet med henblik på at bestemme diagnoser for patienter i patientkohorten. Tabellen DIAGNOSER_UKATEGORISERET indeholder kun ukategoriseret information om patientens diagnose mens DIAGNOSER_KATEGORISERET indeholder en angivelse af patientens hoved- eller bi-diagnose.

ICD-10	Diagnoser	Antal kvinder	Antal mænd
F 10-19	Sindslidelser som følge af misbrug	80	200
F 20-29	Skizofreni og beslægtede diagnoser	96	265
F 30-39	Affektive lidelser	49	61

Tabel 6.5: Patientfordeling for lidelser med indikation for antipsykotika[7]

Da patientkohorten alene skal bestå af patienter med tilfælde af antipsykotisk monoterapi i patientens *første* indlæggelse, og de implicerede behandlingsforløb alene skal udgøres af enten *første* behandlingsforløb eller indlæggelsens *sidste* behandlingsforløb, optræder en yderligere indsnævring af det empiriske materiale. På baggrund af tabeloprettelserne er det muligt at bestemme 130 behandlingsforløb[5] i antipsykotisk monoterapi, der fører til behandlingsskifte og 224 behandlingsforløb[6], der fører til udskrivning. Da det enkelte behandlingsforløb alene vedrører én patient, er der således tale om en patientkohorte på 354 patienter eller et udsnit på knap 18% af patientpopulationen fra den valgte periode.

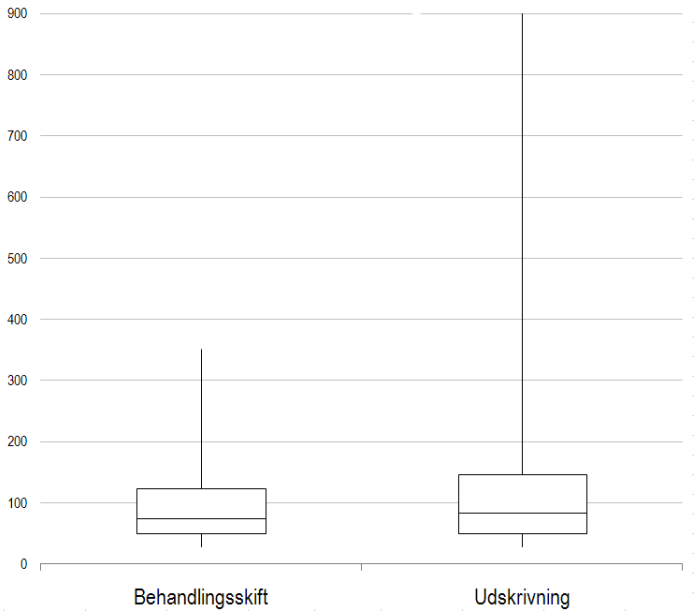
Behandlingsforløb for patientkohorten

For behandlingsforløb omtalt som *patientkohortens behandlingsforløb* er der alene tale om forløb, der enten (1) leder frem til udskrivningen fra patientens første indlæggelse eller (2) leder frem til første indlæggelses første behandlingsskifte. Disse to grupper repræsenterende forsøgets udfaldsparametre vil i det følgende blive omtalt som henholdsvis *behandlingsskiftegruppen* og *udskrivningsgruppen*.

Behandlingsforløbets varighed. Som figur 6.14 illustrerer, er der ikke signifikant forskel på varigheden (på figuren angivet i dage) for de to grupper. For behandlingsskiftegruppen er der tale om en gennemsnitlig varighed på knap 100 dage, imens den for

[5]Se bilag A.3.1
[6]Se bilag A.3.2

udskrivnings-gruppen er på 125 dage[8].

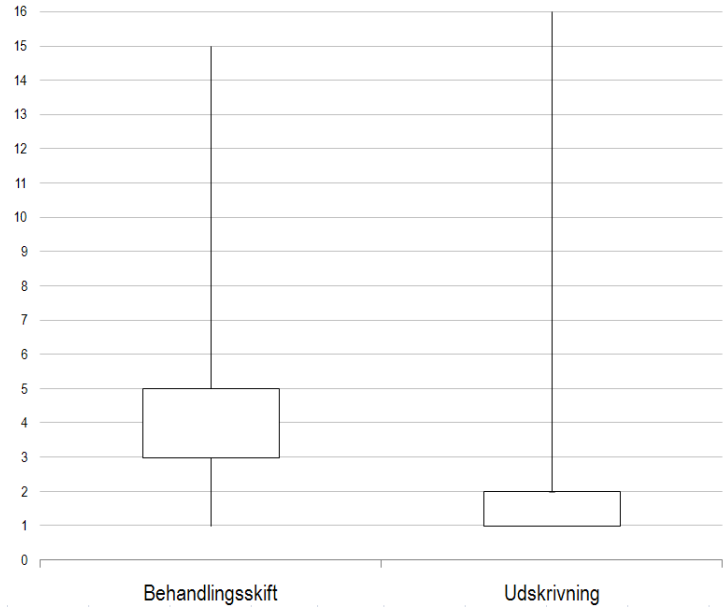


Figur 6.14: Varigheden af behandlingsforløbet udtrykt i dage (y-aksen) for grupperne behandlingsskifte (til venstre) og udskrivning (til højre). Antallet af behandlingsforløb der er udslagsgivende for fordelingen er udtrykt ved et såkaldt *boxplot*: Et boxplot er en figur der opdeler datasættet i *kvartiler* og som benyttes til at beskrive datasættes spredning. Den består af en rektangulær figur gennemskåret af en vandret linie (medianen) og med en stav i forlængelse af figurens top og bund. Den rektangulære figur angiver intervallet for 50% af datasættet, mens medianen (også omtalt som 50%-kvartilen) angiver det punkt på interval-aksen (y-aksen) der deler det samlede datasæt i to lige store dele. For behandlingsskiftegruppen er medianen 75 mens den for udskrivningsgruppen er 84. Bunden af den rektangulære figur udtrykker 25%-kvartilen (værdier herunder udgør 25% af datasættet) mens figurens top udtrykker 75%-kvartilen (værdier herover udgør de resterende 25% af datasættet). Den øvre og nedre stav indikerer dataområdet for disse resterende 50% af datasættet. 25%-kvartilen og 75%-kvartilen udgør for behandlingsskifte-gruppen henholdsvis 49 og 123 mens de for udskrivnings-gruppen udgøres af henholdsvis 50 og 148.

Antal behandlingsforløb i samme indlæggelse. Selvom behandlingsforløbene for de to grupper er udvalgt som henholdsvis første behandlingsforløb (behandlingsskiftegruppen) og behandlingsforløb, der leder frem til udskrivning (udskrivningsgruppen), vil behandlingsforløbene for disse grupper ikke udgøre den samlede indlæggelse og der vil for begge grupper optræde andre behandlingsforløb i løbet af den samme indlæggelse. På figur 6.15 er antallet af behandlingsforløb ved gruppernes samlede indlæggelse illustreret. Som figuren illustrerer, er der en klar tendens til færre behandlingsforløb for patienter i gruppen udskrivning med et gennemsnit på knap 2 behandlingsforløb mod behandlingsskifte-gruppens gennemsnit på 4,2 forløb[9].

Antal journalnotater. Antallet af journalnotater varierer for de udvalgte behandlingsforløb. I specialets undersøgelser tages udgangspunkt i henholdsvis et udsnit af

[8]Se bilag A.3.3
[9]Se bilag A.3.4



Figur 6.15: Antallet af behandlingsforløb indeholdt i hver af gruppernes samlede indlæggelser. På figuren er [25%-kvartilen, medianen, 75%-kvartilen] for behandlingsskiftegruppen udtrykt ved [3, 3, 5], imens den for udskrivningsgruppen er [1, 1, 2]. For bestemmelse af disse værdier henvises til figurteksten for figur 6.14.

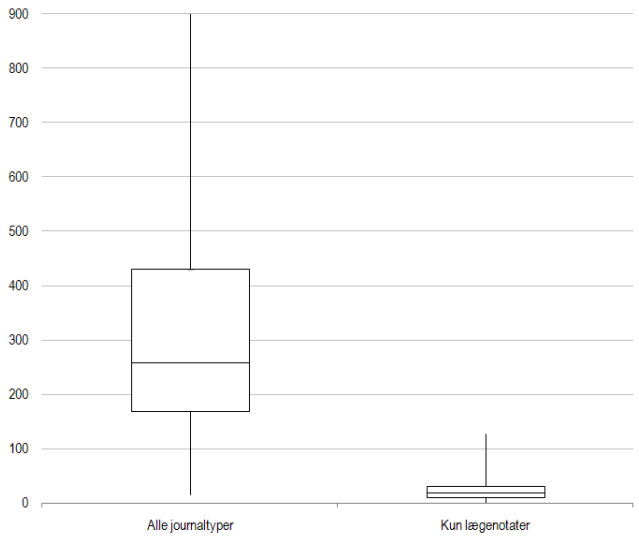
alle journaltyper, hvor der ikke skelnes mellem notatets forfattere og et udsnit af *lægenotater*; det vil sige notater der alene er forfattet af læger (uanset funktion og titel) jf. afsnit 4.4.
Figur 6.16 illustrerer antallet af journalnotater for behandlingsforløbene frem til behandlingsskiftet, imens figur 6.17 illustrerer antallet frem til udskrivning. I begge tilfælde er der tale om væsentlig færre notater, når det udelukkende er notattypen lægenotater, der betragtes. Ved sammenligning mellem behandlingsskiftegruppen og udskrivningsgruppen er forskellen dog minimal. Denne observation stemmer overens med behandlingsvarigheden for de to grupper, hvor varigheden anført i dage stort set er identisk.

Kliniske variable for patientkohorten

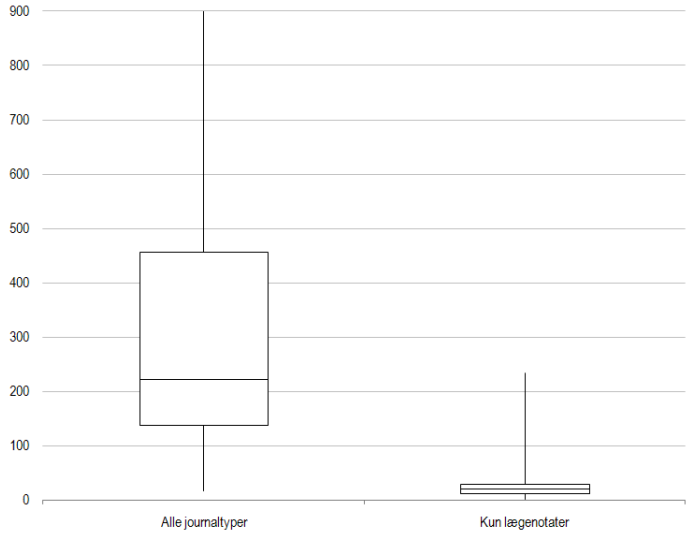
Køn. Figur 6.18 illustrerer kønsfordelingen i grupperne, hvor behandlingsskiftegruppen består af forholdsvis flere kvinder end tilfældet er for udskrivningsgruppen med 70% mod 51%[10].

Alder og afdelingstilknytning. Figurerne 6.19 og 6.20 illustrerer aldersspredningen i henholdsvis behandlingsskifte- og udskrivnings-gruppen. Den væsentligste forskel på aldersfordelingerne optræder i behandlingsforløbene frem til behandlingsskiftet, som illustreret på figur 6.19. Hvor afdelingerne M, L og R fremtræder med stor

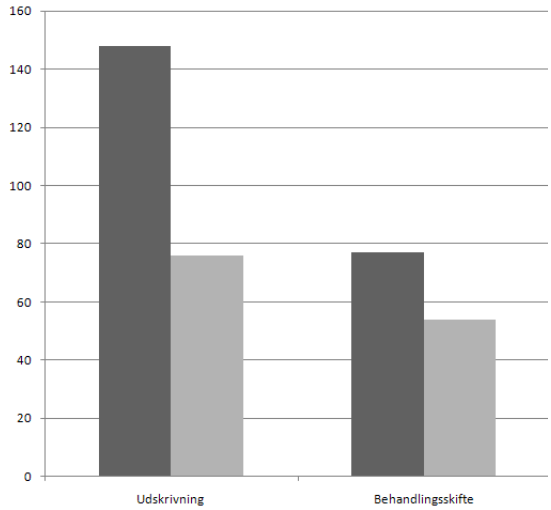
[10]Se bilag A.3.5



Figur 6.16: Antal journalnotater for behandlingsskiftegruppen. *Alle journaltyper* er angivet til venstre på figuren, imens *lægenotaterne* er angivet til højre. På figuren er [25%-kvartilen, medianen, 75%-kvartilen] for alle journaltyper udtrykt ved [170, 258, 431] og for lægenotaterne alene udtrykt ved [11, 20, 31]. For bestemmelse af disse værdier henvises til figurteksten for figur 6.14.



Figur 6.17: Antal journalnotater for udskrivningsgruppen. *Alle journaltyper* er angivet til venstre på figuren, imens *lægenotaterne* er angivet til højre. På figuren er [25%-kvartilen, medianen, 75%-kvartilen] for alle journaltyper udtrykt ved [137, 221, 453] og for lægenotaterne alene udtrykt ved [13, 20, 29]. For bestemmelse af disse henvises til figurteksten for figur 6.14.



Figur 6.18: Kønsfordeling i grupperne med udskrivning i venstre side og behandlingsskifte i højre side. De mørke søjler angiver antallet af mænd, imens de lyse angiver antallet af kvinder. I udskrivningsgruppen er kønsfordelingen (148/76), imens den for behandlingsskiftegruppen er (77/54).

lighed, er patienter på afdeling U væsentligt yngre end gennemsnittet på førnøvnte afdelinger, imens patienterne på afdeling P er væsentligt ældre. Disse forskelle er dog stort set udlignet ved udskrivning jævnfør figur 6.20. Ved sammenligning af de to grupper er aldersspredningen dog stort set identisk som vist på figur 6.21.

Diagnoser. Figur 6.22 angiver antal forekomster af diagnoserne F1-3 ved henholdsvis behandlingsskift og udskrivning. Som figuren angiver, er der for begge grupper flest forekomster af F2 diagnoser efterfulgt af F1 og slutteligt F3 diagnoserne.

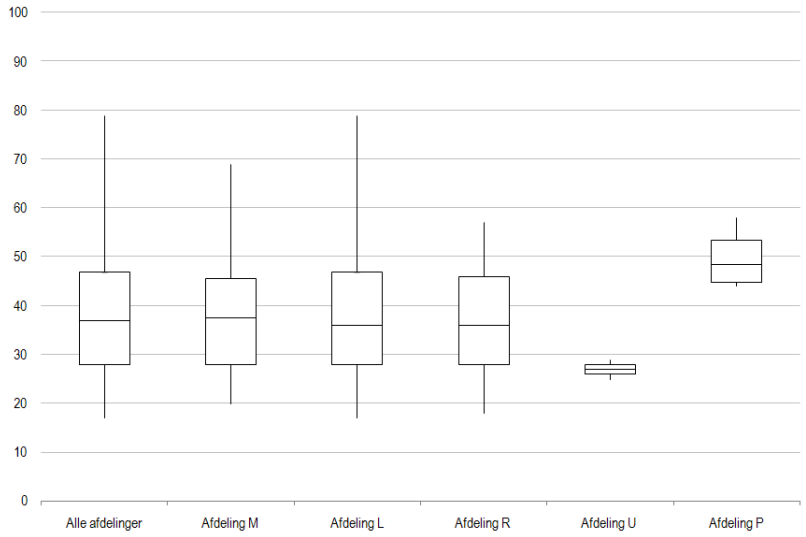
6.4 Parameterselektion

På baggrund af pilotforsøgets tabeloprettelser og identifikation af kandidater til specialets undersøgelser, er det muligt at foretage en klar parameterselektion i EPJ-systemets database.
Denne parameterselektion er gennemgået i det følgende og vedrører:

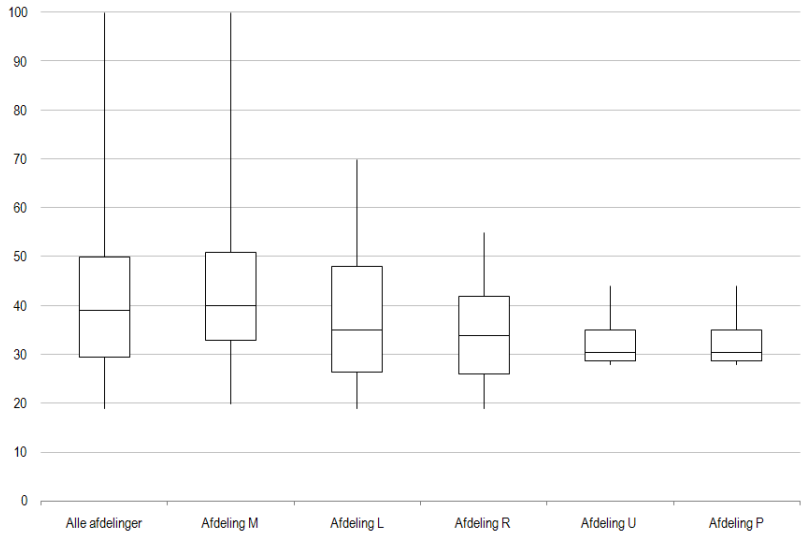
- Dokumentudvælgelse til tekst data mining
- Selektion af kliniske variable
- Selektion af koblede parametre

6.4.1 Dokumentudvælgelse til tekst data mining

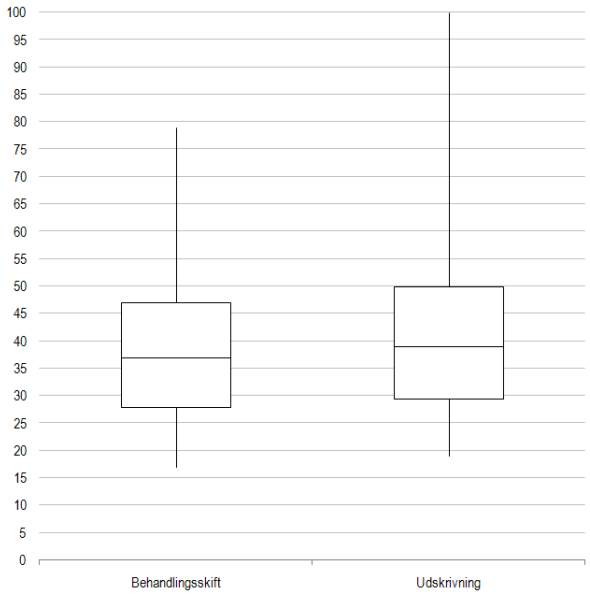
Udgangspunktet for clustering af tekstuelle informationer er et datasæt bestående af elementer i form af såkaldte dokumenter, jf. afsnit 5.4.1. Parameterselektionen



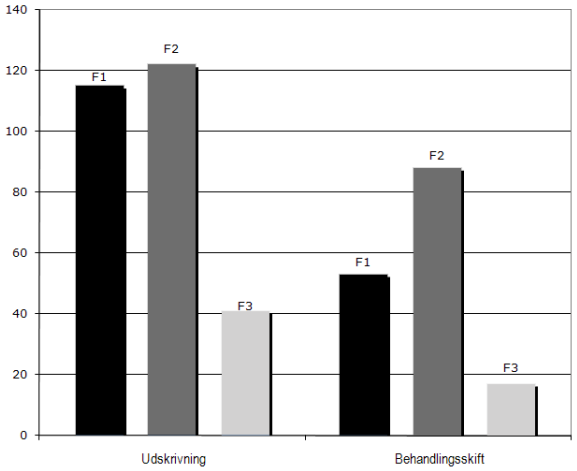
Figur 6.19: Aldersfordeling på de forskellige afdelinger i behandlingsforløb frem til behandlingsskifte. Fra venstre mod højre er det først alderssprednigen for alle afdelinger, derpå er det afdelingerne M, L, R, U og slutteligt afdeling P. På figuren er [25%-kvartilen, medianen, 75%-kvartilen] for alle afdelingerne (130 patienter) udtrykt ved [28, 37, 47] mens den for de enkelte afdelinger er (56 patienter på afdeling M) [28, 38, 46], (57 patienter på afdeling L) [28, 36, 47], (11 patienter på afdeling R) [28, 36, 46], (2 patienter på afdeling U) [26, 27, 28] og (4 patienter på afdeling P) [45, 49, 54]. For bestemmelse af disse værdier henvises til figurteksten for figur 6.14.



Figur 6.20: Aldersfordeling på de forskellige afdelinger frem til udskrivningen. Fra venstre mod højre er det først alderssprednigen for alle afdelinger, derpå er det afdelingerne M, L, R, U og slutteligt afdeling P. På figuren er [25%-kvartilen, medianen, 75%-kvartilen] for alle afdelingerne (224 patienter) udtrykt ved [30, 39, 50], imens den for de enkelte afdelinger er (132 patienter på afdeling M) [33, 40, 51], (63 patienter på afdeling L) [27, 35, 48], (21 patienter på afdeling R) [26, 34, 42], (4 patienter på afdeling U) [29, 31, 35] og (4 patienter på afdeling P) [29, 31, 35]. For bestemmelse af disse værdier henvises til figurteksten for figur 6.14.



Figur 6.21: Aldersspredning for samtlige afdelinger ved henholdsvis behandlingsskift (figuren til venstre) og udskrivning (figuren til højre). På figuren er [25%-kvartilen, medianen, 75%-kvartilen] for behandlingsskiftegruppen (130 patienter) udtrykt ved [28, 37, 47], imens udskrivningsgruppen (224 patienter) er udtrykt ved [30, 39, 50]. For bestemmelse af disse værdier henvises til figurteksten for figur 6.14.



Figur 6.22: Diagnosernes fordeling i de to grupper. Optælling er baseret på ukategoriseret diagnosedata, det vil sige, at der ikke skelnes mellem hoved- eller bidiagnoser. I venstre side er det udskrivningsgruppen (↑), imens den højre side angiver behandlingsskiftegruppen (↓). Tallene i parentes henviser til antallet af patienter. Den mørkeste søjle angiver F1 diagnoser (115 ↑ / 53 ↓), den mørkegrå angiver F2 diagnoserne (122 ↑ / 88 ↓), imens den lysegrå angiver F3 diagnoserne (41 ↑ / 17 ↓).

for tekst data mining processen har derfor til formål af fastlægge indholdet af disse dokumenter.
Følgende selektionssnit af journalnotater udgør parametre til undersøgelse ved clustering:

- Samtlige notater fra behandlingsforløbets start til behandlingsforløbets ophør
- De sidste 10 notater frem til behandlingsforløbets ophør
- De sidste 10% af notaterne frem til behandlingsforløbets ophør
- Samtlige lægenotater fra behandlingsforløbets start til behandlingsforløbets ophør
- De sidste 10 lægenotater frem til behandlingsforløbets ophør
- De sidste 10% af lægenotaterne frem til behandlingsforløbets ophør

Ved undersøgelse af den pågældende parameter udvælges alene datasæt bestående af dokumenter indeholdende det valgte selektionssnit af journalnotater. Journalnotater betegner samtlige notatforekomster i patientjournalen jævnfør afsnit 4.4, hvor lægenotater alene omfatter de notater, der er udfærdiget af læger. I forlængelse af gennemgangen i afsnit 4.4 drejer det sig om følgende notattyper: *forundersøgelser*, *indlæggelsesnotater*, *behandlingsplaner*, *medicinafvigelser*, *behandlingsnotater* og *udskrivelsesnotater*. I enkelte tilfælde vil de endvidere omfatte *indlæggelsessamtaler*, *rehabiliteringsplaner*, *særlige aftaler* samt *epikriser*.
De forskellige selektionssnit af journalnotater er udvalgt med henblik på at bestemme den selektion, der fremkommer med en stærk overensstemmelse mellem den forventede og observerede fordeling. Reduktionen i selektionssnittets størrelse fra *alle* over *de sidste 10%* til de sidste 10 notater forventes at reducere den støj, der vedrører registreringer uden fokus på det foranstående udfald. Valg af lægenotater forventes endvidere at tydeliggøre rationalet for den valgte behandling og dermed at åbenbare den systematik, der forventes at kendetegne lægenotater frem til et givent udfald.

6.4.2 Selektion af kliniske variable

Clusterundersøgelsen af de kliniske variable tager udgangspunkt i et datasæt bestående af den valgte parameter. Jeg har udvalgt følgende parametre til specialets undersøgelser:

- Køn
- Alder
- Afdeling
- Diagnose (kategoriseret som hoved- og/eller bi-diagnose)
- Diagnose (ukategoriseret)

For alle kliniske variable gælder, at de ved dataselektion udtrækkes fra afledte tabeller fremkommet på baggrund af pilotforsøget. De valgte variable er alle relateret til datoen for det pågældende behandlingsforløbs start. Selvom variable som udgangspunkt er konstante for hele perioden fra behandlingsforløbets start til dets ophør, hænger fastlæggelsen af den tidslige ramme sammen med variablen *alder*.
De kliniske variable er udvalgt ud fra hypotesen om, at de udgør parametre for behandlingsudfaldene udskrivning og behandlingsskifte, som begrundet i afsnit 2.2.3. Forventningen er derfor at der blandt de valgte kliniske variable vil optræde markører for førnævnte behandlingsudfald.

6.4.3 Selektion af koblede parametre

Selektionen baseres på enkeltvis udvælgelse af ovennævnte parametre. Først i datatransformationen intergreres de forskellige parametre i samme similaritetstabel til udførsel af clustermetoden.
Udgangspunktet for parametervalget er den samtidige uafhængige tilstedeværelse af flere parametre er markører for behandlingsudfaldene udskrivning og behandlingsskifte. At parametrene betragtes som uafhængige sker for at fastlægge et beregningsmæssigt udgangspunkt for similaritetsbestemmelsen. Valget er naturligvis diskutabelt men er gjort for at undgå kunstigt højere similariteter.
Forventningen til clusterfordelingerne af de koblede variable afhænger i al væsentlighed af udfaldet af de foregående analyser. Med forventning om forekomst af markører blandt både kliniske variable og selektionssnit af journalnotater vurderes de koblede variable at fremkomme med rimelige fordelinger.
De valgte koblede variable omfatter:

- Køn-afdeling
- Køn-alder
- Alder-afdeling
- Køn-alder-afdeling
- Køn-alder-diagnose (ukat.)
- Køn-alder-diagnose (kat.)
- Alder-afdeling-diagnose (ukat.)
- Alder-afdeling-diagnose (kat.)
- Køn-afdeling-diagnose (ukat.)
- Køn-afdeling-diagnose (kat.)
- Køn-alder-afdeling-diagnose (ukat.)

- Køn-alder-afdeling-diagnose (kat.)
- Alder-afdeling-journalnotater

For sidstnævnte kobling gælder at det valgte selektionssnit af journalnotater udgøres af de sidste 10% af *alle journaltyper* med termfrekvens-inversdokumentfrekvens (tf-idf) vægtning.

6.5 Data mining processen

Data mining processen i specialets undersøgelser er opdelt i henholdsvis (1) tekst data mining, (2) data mining af kliniske variable og (3) koblingen mellem tekst data mining og data mining af kliniske variable, hvor de valgte parametre udgøres af henholdsvis journalnotater og tilstedeværelsen af kliniske variable udvalgt i afsnit 6.4.

6.5.1 Tekst data mining

På baggrund af den valgte parameter gennemføres tekst data mining processen som beskrevet i afsnit 5.4.

Dataforbehandling

Ved dataforbehandling med inddragelse af lemmatisering anvendes retskrivningsordbogen der indeholder i alt 237.572 grundstammer. Lemmatiseringstabellen, der oprettes til dette formål, tilføjes desuden *præparatnavne* og *domænespecifikke psykiatritermer*. Præparatnavne udgøres af de stavemåder for antipsykotika, der forekommer i EPJ-systemet, imens de domænespecifikke psykiatritermer er identificeret på baggrund af en indledende *uddragning af domænespecifikke termer*[11] på en række tilfældigt udvalgte journalnotater fra behandlingsforløb på tværs af patientmaterialet i EPJ-systemet. Dette bringer lemmatiseringstabellen op på 237.953 kendte termer.
Lemmatiseringstabellen indeholder som udgangspunkt 2.553 markerede stopord. Ved anvendelse af lemmatisering på datasættets termer foretages derfor en samtidig fjernelse af disse ordforekomster. Lemmatiseringstabellen tilføjes yderligere navneliste for danske navne stillet til rådighed af familiestyrelsen inddragende 18.275 fornavne, der ligeledes tildeles stopords-status.
Som eksempel på lemmatisering betragtes følgende journalnotat:

Pt ringede kl. 13:30 og fortalte at hun ikke havde det så godt efter en bytur, hvor der havde været for mange mennesker. På forespørgsel sagde pt att hun ikke var psykotisk, men gerne ville have noget beroligende. Pt har ikke pn med på weekend, da hun ikke ønskede dette. Pts mor har tbl. rivotril a 0.5 mg liggende og pt spørger om hun må tage en af dem hvorefter hun vil lægge sig og slappe af. Ut siger at det er ok. og underretter vagt-havende om dette.

Efter lemmatisering fremstår det foregående journalnotat på følgende form:

[11] Se afsnit 5.4.4.

ringe fortælle ikke have godt hvor have være mange menneske. forespørgsel sige ikke være psykotisk gerne ville have beroligende. have ikke weekend ikke ønske. mor have liggende spørger må tage hvorefter ville lægge slappe. sige være. underrette vagt havende.

Selvom notatet umiddelbart fremstår væsentligt reduceret, synes hovedbudskabet dog at være bevaret.
Ved *uddragning af domænespecifikke termer* benyttes lemmatiseringstabellen til at frasortere identificerbare grundstammer. Domænespecifikke tilføjelser som præparatnavne og psykiatritermer inkluderes ikke i denne grundstamme-identifikation.

Datatransformation

Ved datatransformation anvendes termvægte og similaritetsbestemmelser, som er beskrevet i afsnit 5.2.3.
Term *boosting* anvendes på baggrund af termer defineret af overlæge Klaus Damgaard Jakobsen (KDJ) fra Psykiatrisk Center Hvidovre. Disse termer er oprindeligt udvalgt ud fra KDJs vurdering af klinikeres termanvendelser i EPJ i forbindelse med dokumentation af behandlingsudfald og er angivet i tabel 6.6.

respons	virkning	effekt
behandlingsrespons	bedring	psykotisk
remission	apsykotisk	responder
psykose	klager	bivirkninger
compliance	bivirkningspræget	eps
dystoni	tardiv	akut
tremor	rigiditet	metabolisk
syndrom	akkomodationsbesvær	akkomodationsvanskeligheder
sukkersyge	diabetes	blodsukker
ortostatisk	hypotension	ostipation
hypersalivation	savlende	lipidprofil
krysteroler	fraktioneret	cholesteroler
ldl	vldl	triglycerider
bmi	prolactin	aseksuel
apsykotisk	impotens	rejsningsbesvær
dysfunktion	seksuel	seksuelle
parkinsonistiske	metabolsk	partiel
behandlingseffekt	produktiv	sygdomsgrad
komplians	akatisi	parkinsonisme
mundtørhed	urinretention	blodtryksfald
forstoppelse	fraktioneret	hdl
vægt	galactose	viagra
ekstrapyramidale	amenore	

Tabel 6.6: Klaus Damgaard Jakobsens vurdering af væsentlige termer anvendt i kliniken. Der forekommer 68 sådanne termer.

Clustering

I alt gennemføres 15 clusterfordelinger på baggrund af k-medoide-algoritmen. Resultaterne herfra er gennemgået i afsnit 7.2.

6.5.2 Data mining af kliniske variable

Idet der er foretaget en række indledende tabeloprettelser i forbindelse med pilotforsøget, kan parametre i form af kliniske variable udtrækkes direkte derfra. Til forberedelse af data mining metoden, har det således alene været nødvendigt at gennemføre trinnet datatransformation[12] efter elementselektionen har fundet sted.

Datatransformation

Ved datatransformationen gennemføres en similaritetsbestemmelse, hvor forekomster af den udvalgte parameter bestemmes ved intervallet [0,1]. Similaritet for parametrene køn og afdeling fortages ved absolut angivelse hvor 0 angiver forskellige køn eller afdelinger, imens 1 angiver samme køn eller afdeling. Similaritet ved de øvrige parametre angives som den procentuelle lighed mellem variable. For de ukategoriserede diagnoser sker det ud fra de tilstedeværende diagnoser, imens det for de kategoriserede diagnoser yderligere tillægges betydning, hvorvidt den samme diagnose også optræder som samme hoved- eller bidiagnose. For variablen alder bestemmes similariteten som den procentuelle andel af den laveste alder i forhold til højeste alder ved parvis sammenligning.
Similaritetsbestemmelsen leder til oprettelsen af et særligt similaritetsmatrice, der benyttes ved data mining metoden.

Clustering

I alt gennemføres fem clusterfordelinger baseret på de fem valgte parametre. Resultaterne herfra er gennemgået i afsnit 7.3.

6.5.3 Data mining af koblede parametre

Data mining af koblede parametre tager udgangspunkt i henholdsvis en kobling mellem forskellige kliniske variable og en kobling mellem bedste parametervalg ved tekst data mining og bedste parametervalg ved koblingen af kliniske variable.

Datatransformation

På baggrund af allerede eksisterende similaritetsmatricer oprettes et similaritetsmatrice, hvor similariteten mellem de valgte parametre udregnes ved at gange værdier for de parvise similariteter med hinanden. Dette valg er baseret på hensynet til sammenlignelighed med undersøgelsens øvrige similaritetsmål, der fremkommer med værdier

[12] Se afsnit 5.2.3.

i intervallet [0,1]. Der er således tale om konjugation mellem similariteterne [Durante *et al.*, 2007]. I litteraturen angives dog andre muligheder så som udtrækning af den fælles laveste værdi eller fælles højeste værdi [Bloch, 2006]. Da mit udgangspunkt dog er at betragte parametrene som uafhængige, har jeg valgt den konjugerende metode.

Clustering

Clustering med k-medoide-algoritmen gennemføres med ialt 13 clusterfordelinger. Resultaterne herfra er gennemgået i afsnit 7.4.

Kapitel 7

Resultater

For samtlige clusterfordelinger har jeg valgt at anvende Kappa-værdier som mål for det observerede udfalds overensstemmelse med den forventede clusterfordeling. Udregning af Kappa-værdien er baseret på en simpel beregning, hvor resultatet antager en værdi i intervallet [-1,1]. Negative værdier angiver et fravær af overensstemmelse, imens værdien 0 er udtryk for en tilfældig fordeling. Positive værdier angiver derimod graden af overensstemmelse mellem det forventede og selve udfaldet, hvor 1 angiver absolut overensstemmelse.

7.1 Kappa-bestemmelse

Kappa-værdien k benyttes blandt andet i klinisk sammenhæng som mål for overensstemmelsen mellem to observationer, hvor hver observation udtrykker en fordeling af samme kliniske materiale [Viera and Garrett, 2005]. Ved alle undersøgelser i specialet har jeg derfor valgt at anvende Kappa-værdier til at beskrive overensstemmelsen mellem den henholdsvis observerede og forventede fordeling.
Udover at være baseret på udregningen af sammenfald i fordelingerne, er Kappa-værdien korrigeret for den overensstemmelse, der kan siges at optræde ved helt tilfældige fordelinger.
Formlerne til beregning af Kappa-værdien er baseret på artiklen *Understanding Interobserver Agreement: The Kappa Statistic* af [Viera and Garrett, 2005] og er gennemgået i det følgende. De variable, der anvendes til beregningen, er hentet fra tabel 7.1. Til udregning af overensstemmelsen P_o mellem den observerede fordeling og den forventede fordeling benyttes formlen (7.1), imens den forventede overensstemmelse P_f mellem fordelingerne fremkommer af (7.2).
Formel (7.3) udregner på baggrund af P_o og P_f Kappa-værdien k for den pågældende fordeling.

$$P_o = \frac{A + D}{NM_{total}} \tag{7.1}$$

	$\mathbf{udfald1}_{frv}$	$\mathbf{udfald2}_{frv}$	total
$\mathbf{udfald1}_{obs}$	A	B	M_1
$\mathbf{udfald2}_{obs}$	C	D	M_2
total	N_1	N_2	NM_{total}

Tabel 7.1: Forholdet mellem den forventede fordeling og den observerede fordeling for de to udfald; udfald1 og udfald2. Tabellen anvendes som baggrund for udregningen af Kappa-værdier for fordelingerne fremkommet ved specialets undersøgelser, hvor (A) beskriver det observerede antal udfald af typen udfald1, der som forventet tilhører udfald1. (B) beskriver derimod de observerede udfald1, der var forventet at tilhøre typen udfald2, ligesom (C) beskriver de observerede udfald af typen udfald2, som var forventet at tilhøre typen udfald1. (D) angiver de udfald med overensstemmelse mellem forventningen og observationen af udfald af typen udfald2. M_1 angiver det totale antal observerede udfald1, M_2 det totale antal observerede udfald2, mens N_1 angiver det totale antal forventede udfald1 og N_2 det totale antal forventede udfald2. (A) og (D) udtrykker således alle tilfælde af overensstemmelse mellem de observerede og de forventede fordelinger, mens (B) og (C) angiver tilfælde af uoverensstemmelse mellem fordelingerne. NM_{total} angiver det totale antal udfald uafhængig af typen.

$$P_f = [(\frac{N_1}{NM_{total}}) * (\frac{M_1}{NM_{total}})] + [(\frac{N_2}{NM_{total}}) * (\frac{M_2}{NM_{total}})] \tag{7.2}$$

$$k = \frac{P_o - P_f}{1 - P_f} \tag{7.3}$$

Som eksempel på udregningen af Kappa-værdier for fordelingerne fremkommet ved specialets undersøgelser, har jeg valgt at tage udgangspunkt i tf-idf vægtningen ved vokabulariet for samtlige journalnotater på baggrund af alle journaltyper (se bilagets tabel C.1). De forventede og observerede fordelinger er angivet i tabel 7.2 og følger opstillingen fra tabel 7.1.
Indsættelse af værdierne fra tabel 7.2 i formlen (7.1), der udregner overensstemmelsen P_o mellem den observerede fordeling og den forventede fordeling, fremkommer med udregningen som angivet på formlen (7.4).
Den forventede overensstemmelse P_f for udfaldene i tabel 7.2 udregnes på baggrund af formlen (7.2) og er angivet i (7.5).
Udregningen af Kappa-værdien k følger nu formlen (7.3) for de fremkomne værdier for P_o og P_f og er angivet i (7.6).

$$P_o = \frac{155 + 53}{354} = 0,587 \tag{7.4}$$

$$P_f = [(\frac{232}{354}) * (\frac{224}{354})] + [(\frac{122}{354}) * (\frac{130}{354})] = 0,541 \tag{7.5}$$

$$k = \frac{0,587 - 0,541}{1 - 0,541} = 0,101 \tag{7.6}$$

Fordi Kappa-værdien er normaliseret ved divisionen med $(1 - P_f)$ jf. formel (7.3), vil k altid falde i det lukkede interval [-1,1]. Værdier for k kan derfor sammenlignes og vurderes i forhold til en fast skala, som udtrykker hvor god overensstemmelsen er

imellem den observerede og den forventede fordeling. En sådan skala, beskrevet af [Viera and Garrett, 2005], er gengivet i tabel 7.3.

	udskrivning$_{frv}$	**behandlingsskift**$_{frv}$	total
udskrivning$_{obs}$	155	69	224
behandlingsskift$_{obs}$	77	53	130
total	232	122	354

Tabel 7.2: Forholdet mellem den forventede fordeling og den observerede fordeling for udfaldende udskrivning og behandlingsskift ved anvendelse af tf-idf vægtningen på vokabulariet for samtlige journalnotater på baggrund af alle journaltyper som parameter. Af de 354 udfald stemmer de 155 forventede udskrivninger overens med den observerede fordeling, ligesom 53 af behandlingsskiftene udtrykker overensstemmelse mellem de observerede og de forventede fordelinger. De 69 og 77 udfald for henholdsvis de forventede behandlingsskift og udskrivninger stemmer således ikke overens med de fremkomne udfald.

Kappa k	**Overensstemmelse**
$0,01 - 0,20$	Svag overensstemmelse
$0,21 - 0,40$	Let overensstemmelse
$0,41 - 0,60$	Moderat overensstemmelse
$0,61 - 0,80$	Stærk overensstemmelse
$0,81 - 0,99$	Meget stærk overensstemmelse

Tabel 7.3: Fortolkning af Kappa-værdien k som beskrevet af [Viera and Garrett, 2005].

Værdien $k = 0,101$, som fremkom ved udregningen i (7.6), ligger indenfor intervallet [0,01,0,20] på tabel 7.3 og angiver derfor en svag overensstemmelse mellem den forventede fordeling og den observerede fordeling for tf-idf vægtningen ved vokabulariet for samtlige journalnotater på baggrund af alle journaltyper.

7.2 Tekst data mining

For at efterprøve hypotesen om tilstedeværelsen af markører for det pågældende behandlingsudfald valgte jeg i første omgang at gennemføre forsøg med dannelse af 2 clusters på baggrund af samtlige parameterangivelser i afsnit 6.4.1. Ved denne forsøgsrække anvendtes ingen metode til dataforbehandling og clusterfordelingerne blev alene dannet på baggrund af termfrekvens-invers dokumentfrekvens (tf-idf) vægtning. De observerede udfald af clusterfordelingerne er fremstillet i bilag C.1.1.

7.2.1 Termfrekvens-invers dokumentfrekvens (tf-idf)

Som det fremgår af tabel 7.4, fremkom det mest signifikante resultat ved Kappa-værdien $k = 0,184$ for udpluk baseret på de sidste 10% af alle journalnotater fra behandlingsforløbet. Ved udpluk baseret på lægenotater var der derimod tale om den ringeste grad af sammenhæng mellem de observerede udfald og de forventede udfald.

Med begrænset tid til gennemførelsen af de videre forsøg blev parameterselektionen forenklet til få repræsentanter for forskellige dataforbehandlings- og termvægtningsmetoder. De to resterende anvendelser af tf-idf vægtningen i undersøgelsen valgte jeg at overføre på alle journaltyper. Af tabel (7.4) ses, at tilføjelsen af både lemmatisering og uddragning af domænespecifikke termer medfører en mindre grad af sammenfald mellem udfaldene og de forventede fordelinger i forhold til anvendelsen af det fulde journalsæt uden lemmatisering. Anvendelsen af lemmatisering medfører størst forringelse med Kappa-værdien $k = 0,027$ mod $k = 0,091$ for uddragning af domænespecifikke termer.

7.2.2 Maksimal termfrekvens (maxtf)

Ved max-tf-vægtning opnåedes den ringeste Kappa-værdi ved alle journaltyper for samtlige notater uden brug af lemmatisering med værdien $k = 0,004$. Til gengæld ændredes værdien væsentligt ved udpluk med de sidste 10 notater frem til behandlingsudfaldet med værdien $k = 0,115$, som samtidig var den næsthøjeste værdi for samtlige clusterfordelinger på journalnotater.

Med tilføjelse af lemmatisering opnåedes generelt gode Kappa-værdier i forhold til de øvrige clusterfordelinger med værdierne $k = 0,082$ og $k = 0,053$ for henholdsvis samtlige journalnotater og de sidste 10 journalnotater frem til udfaldet. På sidstnævnte udpluk valgte jeg at forsøge med *boosting* af kliniske termer, hvilket dog medførte et udfald med den næstlaveste værdi for fordelingerne med værdien $k = 0,005$.

Uddragning af domænespecifikke termer på baggrund af max-tf-vægtning resulterede i værdien $k = 0{,}074$ altså en lavere værdi end tilfældet var med tf-idf-vægtning.

7.2.3 Jaccards similaritetskoefficient

Afslutningsvis valgte jeg at anvende Jaccards similaritetskoefficient, hvilket resulterede i værdien $k = 0,068$, det vil sige en værdi tæt på resultatet af uddragningen af domænespecifikke termer på baggrund af max-tf vægtningen ($k = 0,074$).

7.3 Data mining af kliniske variable

Parameterselektionen baseret på kliniske variable bestod af henholdsvis alder, køn, afdeling og diagnose samt en yderligere opdeling af diagnose efter opdelingen i hoved.- og bidiagnoser. Som vist på tabel 7.5 opnåedes den højeste Kappa-værdi ved clustering baseret på afdelingsnavnet med værdien $k = 0,162$. Kappa-værdien for ukategoriserede diagnoser nærmede sig gennemsnittet for de kliniske variable med værdien $k = 0,081$, imens fordelingen baseret på køn opnåede den noget højere værdi $k = 0,106$. Kategoriserede diagnoser og fordelingen baseret på alder medførte de laveste værdier med henholdsvis $k = 0,024$ og $k = 0,039$.

	Alle journaltyper			Kun lægenotater		
	Alle	Sidste 10	Sidste 10%	Alle	Sidste 10	Sidste 10%
Tf-idf						
Uden lemma	0,101	0,014	0,183	0,004	0,004	0,004
Med lemma	0,027	-	-	-	-	-
Dst.	0,091	-	-	-	-	-
Max-tf						
Uden lemma	0,004	0,115	-	-	-	-
Med lemma	0,082	0,053	-	-	-	-
Klinisk *boost*	-	0,005	-	-	-	-
Dst.	0,074	-	-	-	-	-
Jaccard	0,068	-	-	-	-	-

Tabel 7.4: Kappa-værdier for clusteropdelinger ved tekst data mining på journalnotater. Dst. betegner uddragning af domænespecifikke termer.

Kliniske variable	
Alder	0,039
Køn	0,106
Afdeling	0,162
Diagnose (ukat.)	0,081
Diagnose (kat.)	0,024

Tabel 7.5: Kappa-værdier for kliniske variable.

7.4 Data mining af koblede parametre

Data mining af koblede parametre medførte en generel reduktion i de fremkomne Kappa-værdier. Som vist på tabel 7.6 opnåedes højeste værdi ved kobling mellem alder og afdeling med $k = 0,039$. Et forsøg med at koble den bedst fremkomne parameter ved tekst data mining (tf-idf-vægtning af de sidste 10%) med de koblede kliniske variable alder og afdeling medførte Kappa-værdien $k = 0,144$.
De resterende koblede parametre baseredes alene på kobling mellem kliniske variable. Den laveste værdi opnåedes ved koblingen mellem alder, afdeling og ukategoriseret diagnose med $k = 0,002$, imens de næsthøjeste værdier fremkom ved kobling mellem køn, alder og ukategoriseret diagnose ($k = 0,106$) både med og uden kobling med afdeling.

	Køn	**Diagnose (ukat.)**	**Diagnose (kat.)**	**Afdeling**	**Tf-idf***
Køn				0,101	
Alder	0,101			0,147	
Køn/ Afdeling		0,106	0,014		
Køn/ Alder/ Afdeling		0,106	0,014		
Køn/ Alder		0,032	0,048		
Alder/ Afdeling	0,101	0,043	0,002		0,144

Tabel 7.6: Kappa-værdier for koblede parametre. * angiver, at der er tale om tf-idf-vægtede journalnotater uden anvendelse af lemmatisering på baggrund af de sidste 10% af samtlige journalnotater frem til behandlingsforløbets udfald.

Kapitel 8

Diskusion

Formålet med specialets undersøgelser var at fastlægge hvilke parametre, der var markører for behandlingsudfaldene *udskrivning* og *behandlingsskifte*. For at fastlægge disse markører opdeltes en patientkohorte i to grupper på baggrund af udfaldene af deres behandlingsforløb. Denne opdeling blev betegnet den *forventede fordeling*. For samme patientkohorte udvalgtes parametrene angivet i tabel 8.1 og patientkohorten blev efterfølgende i løbet af undersøgelserne opdelt i to grupper på baggrund af disse parametre. Disse opdelinger blev betegnet de *observerede fordelinger*. Overensstemmelse mellem den forventede fordeling og undersøgelsernes observerede fordelinger blev fastsat på baggrund af Kappa-værdien for fordelingerne. Udgangspunktet for fastlæggelselsen af markører blandt de valgte parametre blev bestemt som Kappa-værdien $k > 0,6$, der udtrykker en stærk overensstemmelse mellem de parvise fordelinger.

8.1 Forsøgsresultater

Specialets undersøgelser dokumenterer ikke en stærk overensstemmelse mellem forsøgsudfald og de forventede udfald ved brug af de valgte parametre fra undersøgelsens datasæt og de valgte analytiske og dataforberedende metoder. Selvom Kappa-værdierne alle var positive og således åbenbarer potentialet for de valgte parametre, opnåede clusterfordelingerne med den stærkeste overensstemmelse kun Kappa-værdier i intervallet 0,01-0,2 (svag overensstemmelse). Kappa-værdierne var samlet set mellem 0,002 og 0,183, hvor den højeste værdi opnåedes ved brug af patienternes journalførte oplysninger i form af *de sidste 10% af journalnotaterne frem til udfaldet ved brug af idf-vægtning til similaritetsbestemmelse*. For de kliniske variable (alder, køn, afdeling og diagnose) opnåedes den højeste værdi ($k = 0,162$) ved brug af afdelingsnavnet som parameter ved similaritetsbestemmelse. Koblingen mellem de forskellige kliniske variable resulterede i Kappa-værdier mellem 0,002 og 0,147 med koblingen *alder og køn* som højeste værdi. Ved kobling mellem den bedste værdi for de journalførte oplysninger og de koblede kliniske variable opnåedes værdien $k = 0,144$.
Kappa-værdierne, der udtrykker forholdet mellem det observerede og det forventede

Kliniske variable
• Alder
• Køn
• Afdeling
• Diagnose (ukat.)
• Diagnose (kat.)[1]
Selektionsnit af journalnotater
Alle journaltyper
• Alle[2]
• Sidste 10[2,3]
• Sidste 10%
Kun lægenotater
• Alle
• Sidste 10
• Sidste 10%
Koblede parametre
• Køn-afdeling
• Køn-alder
• Alder-afdeling
• Køn-alder-afdeling
• Køn-alder-diagnose (ukat.)
• Køn-alder-diagnose (kat.)
• Alder-afdeling-diagnose (ukat.)
• Alder-afdeling-diagnose (kat.)
• Køn-afdeling-diagnose (ukat.)
• Køn-afdeling-diagnose (kat.)
• Køn-alder-afdeling-diagnose (ukat.)
• Køn-alder-afdeling-diagnose (kat.)
• Alder-afdeling-journalnotater[4]

Tabel 8.1: Parametre anvendt i specialets undersøgelser: Bortset fra [1] er de angivne **kliniske variable** anvendt i undersøgelser af hele patientkohorten. For **selektionsnit af journalnotater** er samtlige undersøgelser sket med termfrekvens-inversdokumentfrekvens (tf-idf) vægtning; [2] er desuden anvendt med (1) tf-idf vægtning med lemmatisering; (2) uddragning af domænespecifikke termer; (3) maksimal termfrekvens (maxtf) vægtning med uddragning af domænespecifikke termer og (4) Jaccards similaritetskoefficient; imens [3] yderligere er anvendt med (1) maxtf vægtning; (2) maxtf vægtning med lemmatisering og (3) *boost* af kliniske termer. For de **koblede parametre** angiver [4] at selektionssnittet af journalnotater udgøres af tf-idf vægtning af de sidste 10% af *alle journaltyper*.

resultat, fremkom med et gennemsnit på 0,066 og SD = 0,051 ud af 33 observationer. Således introduceredes en forbedring i fordelingen ved brug af henholdsvis (1) *de sidste 10% af journalnotaterne ved idf-vægtning*; (2) *afdelingsnavnet*; (3) koblingen *alder og køn* og (4) koblingen *alder, køn og de sidste 10% af journalnotaterne ved idf-vægtning*. Resultatet fra specialets undersøgelser indikerer således et særligt forhold ved disse fire potentielt brugbare parametre. Yderligere undersøgelser er dog nødvendige for at fastlægge en egentlig prædiktiv værdi.
I tilfældet med de sidste 10% af journalnotaterne ved idf-vægtning kan det undre, at det netop er denne, der skiller sig væsentligt ud i forhold til de øvrige 14 notat-analyser (den næsthøjeste Kappa-værdi opnåes ved *de sidste 10 notater ved brug af maxtf-vægtning til similaritetsbestemmelse* med $k = 0,115$). Fremtidige undersøgelser, der inddrager de sidste 10% notater, vil således kunne afdække, om der er tale om en generel tendens, eller det blot er udtryk for en tilfældighed ved brug af den valgte metode.
Støj er dog formentlig en væsentlig årsag til, at clusteranalyser baseret på samme metodetilgang fremkommer med overraskende forskellige resultater. Eksempelvis fremkom analyser med tf-idf-vægtede termer med noget forskellige Kappa-værdier, når antallet af journalnotater som eneste variabel ændredes ($k = 0,101$ ved alle notater, $k = 0,014$ ved de sidste 10 notater og $k = 0,183$ ved de sidste 10% af journalnotaterne).
At lemmatisering ikke bidrog til en væsentlig forbedring af clusterfordelingerne hverken ved tf-idf vægtning eller maxtf vægtning var stor overraskelse. Jeg havde som udgangspunkt valgt at tilføje en lang række kliniske termer til den anvendte lemmatiseringstabel i den overbevisning, at fordelingen således ville blive styrket betydeligt hvilket imidlertid ikke var tilfældet.
Et andet overraskende udfald fremkom ved anvendelsen af kliniske termer som parameter for tf-idf-vægtningen. Uanset udsnittet blev udfaldet af clusterfordelingerne det samme ($k = 0,004$ ved alle notater, $k = 0,004$ ved de sidste 10 notater og $k = 0,004$ ved de sidste 10% af lægenotaterne). Lægenotater udgør en relativ lille delmængde af den samlede mængde journalnotater og repræsenterer derfor et oplagt mål for en reduktion af vokabulariet for undersøgelsen. Da disse yderligere forventedes at indeholde de væsentligste eksplicitte rationaler ved behandlingen, var forventningen, til anvendelsen af dette udsnit derfor også, at det ville føre til en endnu stærkere overensstemmelse mellem fordelingerne. At fordelingerne på baggrund af dette udsnit fremkom med så ringe overensstemmelse er derfor overraskende. Til forskel fra de øvrige journalnotater er lægenotaterne præget af en repeterende gennemgang af baggrunden for indlæggelsen, symptombilledet og den planlagte behandling. Der er således en ekstrem lille variation i vokabulariet uafhængigt af det valgte snit. Selvom dette formentlig er en medvirkende årsag til fraværet af variation i clusterfordelingerne for de forskellige udsnit med lægenotater, forklarer det ikke, hvorfor fordelingerne fremtræder ringere end ved anvendelse af alle journaltyper. Den ringe værdi blev yderligere bekræftet i undersøgelsen af term *boosting*, hvor særligt udvalgte kliniske termer fik en kunstig højere vægtning end øvrige termer og fremkom med resultatet $k = 0,005$.

For analyserne med journalnotater som anvendt parameter gælder således, at der generelt fremkommer højere Kappa-værdier ved undersøgelser baseret på dokumenter med forholdsmæssigt mange journalnotater. Forventningen om en negativ sammenhæng mellem størrelsen af vokabulariet og Kappa-værdien blev således ikke opfyldt. Årsagen hertil er muligvis, at et større vokabularium er nødvendigt for at sikre systematikken på tværs af selektionssnittet af journalnotater. Forskellen mellem de forskellige fordelinger er dog minimale, hvorfor det ikke er muligt at konkludere entydigt på årsagen til fordelingerne. En udredning af dette forhold synes at fordre indsamlingen af journalnotater fra en langt større patientkohorte. En sådan ville imidlertid bryde med det valgte forsøgsdesign og er derfor ikke forsøgt gennemført.

Som supplement til det anvendte smalle vindue *frem til* forsøgsudfaldet kunne det dog have været interessant også at betragte journalnotaterne ved indtræden i den pågældende behandling. Eksempelvis de 10 *første* journalnotater i stedet for de 10 sidste. Et sådant udsnit kunne være med til afdække, om et sådan udsnit vil medføre en større systematik end journalnotaterne, der optræder forholdsvist tæt på udfaldet.

At de valgte kliniske variable ikke fremkom med en bedre fordeling er mindre overraskende. De to valgte udfaldsparametre begrænser udfaldsrummet væsentligt, og det er da også parameteren *afdeling*, der som udgangspunkt er baseret på få klare *baggrundsvariable* (afdeling L, M, R, U og P), som klarer sig bedst ($k = 0,162$). Med baggrundsvariable menes at parameteren i sig selv kan siges at udgøres af variable, der bestemmer dens værdi eller i dette tilfælde parameterens navn. Dette stemmer overens med observationen af, at de ukategoriserede diagnoser opnår en bedre fordeling ($k = 0,081$) end de kategoriserede diagnoser ($k = 0,024$), hvor førstnævnte alene er baseret på tre mulige diagnoser (F1, F2 og F3), imens sidstnævnte yderligere er opdelt i hoved- og bidiagnoser med mulighed for flere samtidige diagnoser. Indtroduktionen af et større udfaldsrum må derfor forventes at give udslag i langt mere interessante fordelinger for de valgte kliniske variable, selvom dette naturligvis vil kræve en ændring af det valgte forsøgsdesign. En mulighed havde været at introducere hierarkisk clustering med det formål, at betragte udfaldende ved indsnævring fra det maksimale antal clusters[1].

Ved koblingen mellem både kliniske variable og journalnotater anvendtes en særdeles restriktiv tilgang, idet udgangspunktet var at betragte disse parametre som uafhængige, hvorved konjuktionen mellem similariteterne blev bestemt ved produktet af de enkelte similariteter. Selvom addition af similariteterne for de valgte parametre formentlig havde medført langt bedre Kappa-værdier for fordelingerne, har jeg i forhold til det valgte forsøgsdesign ikke fundet belæg for et sådan valg. Der er således behov for en væsentlig forbedring af Kappa-værdierne for de enkeltvise parametre før resultatet af de koblede parameter vil være af nævneværdig interesse.

[1] Se afsnit 5.2.4 for en beskrivelse af den hierarkiske clustering-metode.

8.2 Undersøgelsesdesign

8.2.1 Inklusionskriterier

Jeg har ved bestemmelse af inklusionskriterierne i afsnit 2.5, haft følgende som mål:

- At opnå så stor en patientpopulation som muligt
- At opnå så komplet beskrevet en patientpopulation som muligt
- At selektere patientforløb med væsentligste behandlingskompetence
- At selektere patientforløb med væsentlig behandlingseffekt af den intentionerede medicinske behandling
- At minimere bias
- At selektere de mest simplificerede patientforløb
- At selektere de patientforløb med størst potentiale for en høj grad af komplians
- At selektere patientforløb med så ensartede medicinprofiler som muligt

Patientgrundlaget udgøres af patienter med afsluttede indlæggelser registreret på sengeafsnit på Psykiatrisk Center Sct. Hans i perioden fra 1. april 2003 til 1. april 2008.
Argumentet herfor er at opnå så stor og velbeskrevet en patientpopulation som muligt. Dette opnår jeg kun ved at gøre rekrutteringsperioden så stor som mulig samtidig med, at der tages højde for at overgangen til EPJ-systemet i indføringsfasen er præget af mindre konsistens i data og fravær af visse registrering såsom dosisstørrelser for antipsykotika-ordinationerne.
Patienter skal have minimum en af følgende diagnoser: F1, F2 eller F3.
Diagnosevalget skal sikre, at den valgte behandling hører under centrets væsentligste behandlingskompetence, hvor klinikerne har størst indsigt i patienternes lidelser, hvilket sikrer de mest pålidelige og målrettede valg i løbet af behandlingen og minimerer støjen i form af tilnærmelsesvis tilfældige behandlingsudfald. De målrettede valg er desuden helt centrale for den valgte farmakologiske behandling og adskiller sig fra de tilfælde hvor antipsykotika benyttes til andre behandlingsformer. Det gælder således generelt, at den viden jeg ønsker at uddrage, skal være så velfunderet som mulig med højest mulige signifikans.
Patienter bidrager med journalførte oplysninger fra behandlingsforløb registreret under patientens første indlæggelse.
For at minimere betydningen og indflydelsen af tidligere indlæggelser på den valgte behandling, har jeg valgt, at der alene skal være tale om patientens første indlæggelse. Dette har således til hensigt at minimere bias. Med bias tænkes her på de erfaringer både klinikere og patient har gjort sig ved tidligere indlæggelser af den pågældende patient.
Behandlingsforløb skal udgøres af minimum 4 ugers antipsykotisk monoterapi.

Ønsket om at vælge patientforløb med så høj en grad af komplians hænger sammen med målsætningen om størst sammenlignelighed samtidig med, at den medicinske behandling har størst potentiale for at influere på behandlingseffekten. Desuden har valg af monoterapi den effekt, at sammenligneligheden mellem de bagvedliggende medicinprofiler øges ligesom behandlingsforløbene vil være relativt simplificerede. Argumentet for at vælge så simplificerede behandlingsforløb og så ensartede medicinprofiler som muligt er at gøre de valgte patienter sammenlignelige samtidig med at undersøgelsens resultater i højere grad vil kunne relateres den øvrige litteratur særligt med hensyn til behandlingseffekter og bivirkninger. Med en varigheden af minimum 4 uger er forudsætningerne desuden til stede for, at behandlingsudfaldet er fremkommet på baggrund af det pågældende behandlingsforløb.
Behandlingsforløb skal føre til enten behandlingsskifte eller udskrivning; for behandlingsforløb førende til behandlingsskifte skal der være tale om indlæggelsens første behandlingsforløb.
Argumentet for dette valg er en klar opdeling af patientkohorten i to grupper, der øger sammenligneligheden i hver af grupper og dermed muliggør fastsættelsen af en *forventning* til opdelingen i undersøgelserne. Dette er endvidere grundlaget for det valgte undersøgelsesdesign.
Patienten bidrager med kun et behandlingsforløb. Hvis patientens indlæggelse kan bidrage med flere behandlingsforløb til undersøgelsen selekteres alene det første behandlingsforløb.
Dette valg er truffet for at sikre, at udfaldet i specialets undersøgelser er fremkommet på så validt et grundlag som muligt og *sampling bias* undgåes.

Selektionssnit

De valgte inklusionskriterier har medført et betydeligt selektionssnit, der har minimeret patientkohortens størrelse betydeligt i forhold til det tilgængelige patientmateriale i EPJ-systemet. Selvom konsekvensen af de stringente inklusionskriterier har medført en stærk reduktion af patientkohorten, repræsenterer det valgte snit et minimum i forhold til opfyldelsen af specialets målsætning. For at forenkle patientforløbene yderligere kunne jeg have introduceret et diagnosekrav alene indenfor F2-spektret for at fremhæve psykoseforekomster. Et væsentligt argument imod dette er imidlertid, at de skizofrene har de værste psykotiske lidelser med relativt få tilfælde af monoterapi. En nærmere undersøgelse alene baseret på denne patientgruppe må nødvendigvis bygge på et væsentligt andet forsøgsdesign end det har været tilfældet i dette speciale.
Inden forsøgenes påbegyndelse valgte jeg at gennemføre et pilotforsøg, der tog udgangspunkt i undersøgelsens inklusionskriterier, der ud over at facilitere den videre data mining proces, gav mig et væsentligt grundlag for at træffe endegyldigt beslutning om fastholdelse af de valgte kriterier. Resultatet af pilotforsøget blev, at ialt 354 behandlingsforløb blev identificeret, hvilket jeg ved sammenligning med andre kliniske undersøgelser finder som værende et fornuftigt patientgrundlag at bygge de videre undersøgelser op omkring.
Det valgte undersøgelsesdesign er primært baseret på en medicinalbiologisk tilgang,

hvilket naturligvis styrker undersøgelsens resultater ud fra både valg af parametre og udfaldsparametre. Ud fra et datalogisk synspunkt kan denne tilgang imidlertid betragtes som særdeles restriktiv, hvor en stor del af den systematik der er nødvendig for at fremkomme med de mest velbegrundede resultater ved clusterundersøgelserne går tabt. Både indskrænkning i form af det empiriske grundlag som de valgte inklusionskriterier medfører og begrænsning af indsamlingsperiode for dokumenter fastlagt ved behandlingsforløb i monoterapi, har således den konsekvens, at det mindsker generaliserbarheden betydeligt af de valgte journalnotater ligesom en stor del af systematikken i de valgte selektionssnit af journalnotater går tabt.

8.2.2 Behandlingsforløb

Behandlingsforløb er i afsnit 2.4.5 defineret som perioden med samme kontinuerlige medicinske behandling. Idet dosisstørrelse ikke medtages på grund af fravær af sådanne registreringer, jf. afsnit 2.6, ligesom det alene er faste ordinationer med antipsykotika, der medregnes som medicinske behandlinger, er der naturligvis tale om et stærkt forenklet syn på den tilstedeværende medicinprofil. En sådan har dog været nødvendig i forsøget på at skabe så sammenlignelige perioder som muligt. Idet den valgte medicinske behandling ikke indgår som parameter i specialets undersøgelser, er det dog min vurdering, at de fastlagte behandlingsforløb styrker undersøgelsernes grundlag.

8.2.3 Pilotforsøg

Til pilotforsøget oprettede jeg en applikation til gennemkørsel af samtlige registrerede præparatordinationer med antipsykotika. Ved første gennemkørsel blev monoterapikandidater identificeret og gemt i en midlertidig tabel med de pågældende patienters løbenumre (PID). Denne gennemkørsel var relativ simpel, idet der ikke blev taget højde for klassificering af flere samtidige identiske ordinationer som monoterapi. Problemet, som dette rejser, er naturligvis, at patienter i forbindelse med flere samtidige ordinationer af samme præparat fejlagtigt identificeres som polyfarmaci-positive på det pågældende tidspunkt. Der er imidlertid flere grunde til, at jeg har valgt dette indledende selektionsgrundlag. For det første vil registreringen af en patient som polyfarmaci-positiv *ikke* forhindre, at patienten på et andet tidspunkt registreres ved sin monoterapi, såfremt der blot er én forekomst af dette. For det andet betragtes centerets patienter over en 5-årig periode, hvor hovedparten af patienterne er gengangere, hvilket øger sandsynligheden for en korrekt identifikation. Slutteligt skal det pointeres, at der til EPJ-systemet knytter sig en meget væsentlig problemstilling: registreringsfejl (støj). Ved at tage udgangspunkt i en så restriktiv selektion øges sandsynligheden for, at patientforløb med forekomster af flere fejlregistreringer sorteres fra.
Pilotforsøget havde yderligere til formål at stå for den indledende kliniske datastrukturering for at facilitere den videre data mining proces og mindske det nødvendige metodeapparat herfor. Dette har ført til oprettelsen af tabeller med behandlinger både

i form af monoterapi og polyfarmaci samt tabeller over indlæggelser på baggrund af grupperede afdelingsindlæggelser. Patientkohorten herfor er baseret på pilotforsøgets patientidentifikationer.
Den valgte fremgangsmåde har været en væsentlig fordel for implementeringen af data mining processen, idet forsøgsdesignet skabte et uundværligt overblik over de tilgængelige data. Desuden er det min vurdering, at den i stor udstrækning har været med til at minimere støj fra datasættet. Støj der ellers skulle være identificeret i datasættet på baggrund af matematiske mål, der uanset pålidelighed ville være medvirkende til en støjreduktion på bekostning af informationerne i datasættet.
Oprettelsen af tabeller på baggrund af det eksisterende datamateriale i EPJ-systemet introducerer naturligvis også støj i form af forsimplinger. Jeg har dog bestræbt mig på at lave en lang række stikprøver i de afledte tabeller for at sikre konsistens mellem disse tabeller og det oprindelige datamateriale. Ud over at facilitere data mining processen har tabeloprettelserne været direkte nødvendige for at kunne gennemføre specialets undersøgelser herunder alle parameter-udvælgelser både i form af kliniske variable som afdeling, alder og diagnose ved behandlingsstart og termforekomster i EPJ-systemet under den pågældende behandling.

8.3 Data mining processen

Data mining processen er som sagt faciliteret væsentligt af det indledende arbejde i specialets pilotforsøg. Overordnet set er den bygget op om følgende processer: (1) En data mining proces med clustering af behandlingsforløb med udgangspunkt i tilstedeværende kliniske variable ved behandlingsforløbets start; (2) en tekst data mining proces med clusterfordeling af behandlingsforløb baseret på journalnotater samt (3)en data mining proces med kobling mellem udvalgte kliniske variable og koblingen mellem de bedste udfald fra hver af grupperne.
Det overordnede formål med disse data mining processer var at bestemme de valgte parametres betydning for clusterfordelingerne. Overensstemmelse mellem de *forventede* fordelinger (der blev fastsat ud fra specialets udfaldsparametre) og de observerede clusterfordelinger blev vurderet ved anvendelse af Kappa-værdier.

Valg af parametre

Valg af parametre (se tabel 8.1) blev truffet på baggrund af det tilgængelige datamateriale og hypotesen om betydningen af disse parametre. Betydningen af de tekstuelle informationer fra journalnotater i EPJ-systemet er desuden begrundet i følgende hypotese[2]:

> at forløbet af indlæggelsen er determinerende for behandlingseffekten og dermed behandlingsudfaldet, må de journalførte informationer frem til behandlingsudfaldets indtræden indeholde markører herfor.

[2]Se afsnit 2.2

Afdeling og diagnoser er medtaget som kliniske variable ud fra en hypotese om, at de to faktorer spiller med i forhold til de valg, der bliver truffet i løbet af patientens behandling på Psykiatrisk Center Sct. Hans. Afdelingerne på Psykiatrisk Center Sct. Hans udtrykker eksempelvis forskellige behandlingsmål, imens forskellige diagnoser potentielt kan afstedkomme forskellige tiltag udtrykt ved den pågældende behandling. Valget af de to sidste kliniske variable køn og alder er udover at være begrundet i normerne ved klinisk epidemiologiske undersøgelser motiveret af sygdomsudtrykket hos den enkelte patient, ligesom de potentielt kan ligge til grund for den valgte behandling.
En anden væsentlig variabel, der *kunne* have været introduceret i specialets undersøgelser er klassifikationen af det til behandlingen anvendte præparat. Idet der er væsentlig forskel på antipsykotikas receptorprofiler, havde det været naturligt at undersøge denne faktors betydning også.

Valg af udfaldsparametre

Valget af udfaldsparametre faldt fra starten på henholdsvis behandlingsskift og udskrivinger, fordi de to er gensidigt ekskluderende og udgør to klinisk set vidt forskellige behandlingsudfald.
Af andre behandlingsudfald, som har været overvejet kan nævnes *indlæggelsestid*, *præparatvalg* og *behandlingsvarighed*. Jeg tog dog udgangspunkt i udfaldende behandlingsskift og udskrivning ud fra hypotesen om, at disse ville udgøre de både mest betydende og mest tydeligt identificerbare udfald.

8.3.1 Støjreduktion

Uanset parametervalg er støjen direkte betinget af den enkelte registrering i EPJ-systemet. Væsentligste forskel er dog brugen af kliniske variable, hvor de valgte datasættet stammede fra afledte tabeller. I forbindelse med pilotforsøgets tabeloprettelser var det alene diagnoserne, der ikke var entydigt identificerbare i EPJ-systemets database. Således identificeredes flere tabeller i EPJ-systemets database med angivelser af diagnose. Ingen af de identificerede tabeller var dog fuldt ud dækkende for hele patientkohorten, og selvom diagnoser identificeredes for alle patienter, var det kun 133 af patienterne, at diagnosen var kategoriseret som patientens bi- eller hoveddiagnose. Dette har medført, at der for anvendelsen af diagnose som klinisk variabel er oprettet to tabeller; den ene med kategorisering af diagnosen, den anden uden. Hvor meget støj dette forhold medfører er imidlertid ukendt. Generelt opnår anvendelsen af de ukategoriserede diagnoser dog højere Kappa-værdier ved clusterfordeling end tilfældet er for de kategoriserede diagnoser.
Til forskel fra diagnoserne kunne de øvrige kliniske variable uden problemer udledes på baggrund af den eksisterende struktur, og de er således ikke mærket af anden støj end den, der skyldes fejlregistreringerne introduceret af EPJ-systemets brugere. Jeg vurderer dog, at denne støj i al væsentlighed koncentrerer sig om de tekstuelle informationer, imens registrering af kliniske variable kun i meget få tilfælde er fejlbe-

hæftede. Uanset valg af parameter har jeg dog ikke foretaget noget indledende træk for at reducere den støj, der skyldes den enkelte registrering.
Der kan i forlængelse heraf argumenteres for, at en indledende støjreduktion i de tekstuelle informationer havde været hensigtsmæssig, men problemet hermed er imidlertid, at kompleksiteten i en sådan opgave er enorm. Alene størrelsen af tekstmaterialet repræsenterer en væsentlig hindring til anvendelse af støjreducerende standardmetoder, der primært er baseret på støjreduktion i nummerisk materiale. For at imødegå den støjproblematik, som anvendelsen af tekstuelle informationer skaber, har jeg valgt at transformere datasættet ved hjælp af flere forskellige typer datatransformation som lemmatisering og uddragning af domænespecifikke termer. Ligeledes har jeg valgt at anvende en række forskellige similaritetsmål for at betragte det støjreducerende potentiale som følge af det enkelte forsøgsudfald.
Støj er formentlig en væsentlig faktor for at clusteranalysen på forskellige udsnit med samme metodetilgang er fremkommet med uforventeligt variende resultater. For eksempel ved anlyse på baggrund af idf-vægtede notater fremkom noget forskellige Kappa-værdier, når antallet af journalnotater som variabel ændredes ($k = 0,101$ ved alle notater, $k = 0,014$ ved de sidste 10 notater og $k = 0,183$ ved de sidste 10% af journalnotaterne).
EPJ-systemets journalnotater indeholder en lang række forkortelser, ligesom fejlstavninger er hyppigt forekommende. Opsætning af dyberegående metoder med semantisk analyse vil næppe have den tilsigtede effekt og støj fremstår derfor - desværre - som et nødvendigt onde ved hovedparten af de registrerede journalnotater.

8.3.2 Metodevalg

Metodevalget omkring data mining processen er begrundet delvist i de metoder, der anvendes som foretrukne metoder i litteraturen særligt med hensyn til undersøgelsen af dokumenter, samt i det forhold at metoderne skulle være så gennemskuelige som muligt for at lette implementeringen af på et på forhånd ukendt materiale og fremme formidlingsaspektet af undersøgelsen.

Vector space modellen

Vector space modellen er grundlæggende for hovedparten af den litteratur jeg har stiftet bekendtskab med i forsøget på at finde relaterede undersøgelser. Jeg har derfor i forlængelse af det forrige valgt, at lade *vector space* modellen danne grundlag for forståelsen af dokumenterne.

Termvægtning

Som termvægte har jeg valgt (1) den maksimale termfrekvens (maxtf); (2) termfrekvens-invers dokumentfrekvens (tf-idf) samt (3) term *boosting*.
De to første er ligesom *vector space* modellen begrundet i litteraturen som helt centrale termvægte. Term *boosting* repræsenterer ikke en samme må velbegrundet metode,

idet der principielt er tale om nogle *biased* valg. Fordelen har imidlertid været, at jeg har været i stand til at belyse udvalgte kliniske termers betydning for undersøgelserne.

Similaritetsmål

Som similaritetsmål har jeg valgt at anvende henholdsvis cosinus og Jaccards similaritetskoefficient. Begge har de tidligere nævnte fordele ved at være beskrevet og samtidig repræsenterer de begge normaliserede similaritetsbestemmelser, hvilket åbner mulighed for at sammenligne de fremkomne fordelinger.

Clustering

Valget af data mining metode faldt på clustering implementeret som k-medoide. Fordelene ved at tage denne metode i anvendelse er flere. For det første er clustering som metode intuitiv forståelig. Formidlingsmæssigt øger dette aspekt således tilgængeligheden af en gren af den medicinske informationsteknologi, der umiddelbart kan fremstå som en relativ kompleks størrelse.
En anden væsentlig pointe er, at metoden er relativ let at implementere, hvilket gør data mining processen lettere at gennemføre på trods af, at jeg har været begrænset i tid til gennemførelse af de enkelte delkomponenter ved data mining processen. Metoden har yderligere den fordel, at den er generaliserbar hvilket gør, at jeg uden videre har kunnet gennemføre forsøgene uden at tage højde for, hvilken parameter, der skulle analyseres ved hjælp af metoden.
Sluttelig er clustering kendetegnet ved at være robust, idet den som metode er mindre følsom overfor støj.
Det sidste argument ændrer dog ikke på, at udfaldet af clusteranalysen kan være influeret af støj, idet clusterantallet ved alle parameteranalyser er fastsat til 2. Jeg har dog i et indledende forsøg afprøvet med clustering i flere grupper dog uden væsentlig generel forbedring i de fremkomne clusterfordelinger af elementerne. Sådanne forsøg svarer til anvendelsen af clustering til støjreduktion til hvilket formål metoden beskrives i litteraturen som anvendelig. Skulle dette have været gennemafprøvet, havde jeg imidlertid måtte inddrage mål for støjselektionen, noget som anvendelsen af hierarkisk clustering af datasættet givetvis kunne have faciliteret. Da jeg som udgangspunkt har vurderet støjen til at være jævnt fordelt over samtlige journaloptegnelser, har jeg imidlertid bevidst fravalgt dette tiltag.

8.4 Fravær af tilsvarende undersøgelser

At jeg ikke har fundet dokumentation på tilsvarende undersøgelser, har naturligvis betydet, at specialets undersøgelser har været karakteriseret ved eksploratorisk udredning af datamaterialet og forsøgsafprøvninger i forskellige retninger for at udvikle en platform at basere undersøgelsen på og relatere dens delkomponenter til. Selvom undersøgelserne ikke er fremkommet med de forventede resultater, er det mit håb, at de vil være en inspiration til fremtidig forskning i kliniske databaser.

Kapitel 9

Konklusion

Problemformuleringen for specialet lød:

> Hvilke markører kan identificeres i databasen for elektroniske patientjournaler (EPJ) på Psykiatrisk Center Sct. Hans som prædikterende for behandlingsudfaldene udskrivning og behandlingsskifte ved psykiatriske diagnoser, hvor behandling med antipsykotika er indiceret?

På baggrund af en patientkohorte på 354 psykiatriske patienter, hvoraf de 130 bidrog med behandlingsforløb førende til behandlingsskifte, imens de resterende 224 bidrog med behandlingsforløb førende til udskrivninger, undersøgtes relationen mellem de observerede fordelinger; de tilstedeværende kliniske variable køn, alder, diagnose og afdelingstilknytning ved behandlingsforløbets start samt selektionssnit fra journalnotater; og de forventede fordelinger med det formål at bestemme de parametre, der førte til en stærk overensstemmelse mellem fordelingerne. Trods gentagne kørsler med clusterfordelinger baseret på forskellige kliniske variable og selektionssnit fra journalnotater opnåedes kun svag overensstemmelse mellem fordelingerne. For de kliniske variable fremkom afdelingstilknytningen med den højeste overensstemmelse bestemt ved Kappa-værdien $k = 0,162$, der på nær en enkelt clusterfordeling opnåede bedre overensstemmelse end samtlige selektionssnit blandt journalnotater. Det bedste resultat opnåedes ved selektionssnit blandt journalnotater med Kappa-værdi $k = 0,183$ ved selektion af de sidste 10% af alle journaltyper frem til udfaldet med tf-idf-vægtning uden anvendelse af dataforberedende metoder som lemmatisering, stop-ords-fjernelse og uddragning af domænespecifikke termer.
De mest overraskende resultater fremkom ved selektionssnit blandt lægenotater ($k = 0,004$) og term boosting blandt kliniske fagtermer ($k = 0,005$), der alle medførte markant dårligere fordelinger end de øvrige selektionssnit blandt journalnotater. Et forhold der på en gang udfordrer antagelsen om potentialet for anvendelsen af de kliniske notattyper og åbenbarer potentialet for anvendelsen af alle journaltyper.
Uddragning af domænespecifikke termer som dataforbehandlende metode fremkom med enten bedre eller næsten samme Kappa-værdi som anvendelsen af lemmatisering.

Selvom der var en forventning om introduktion af mere støj ved førstnævnte metode, viser specialets undersøgelser således, at dette ikke var tilfældet.
At anvendelsen af Jaccards koefficient og cosinus med forskellige termvægtningstiltag ikke fremkommer med væsentlig forskellige resultater indikerer et behov for revision af enten forsøgsdesign eller delvis strukturering af journaldata før identifikation af markører for behandlingsudfald vil være forventelig ved selektionssnit af journalnotater.
Der var ikke nævneværdig forskel imellem Kappa-værdier for de kliniske variable alder, køn, afdeling og diagnose. Dog var der en tendens til lavere Kappa-værdier for parametre baseret på flere baggrundsvariable, hvilket anses som en indikation af nødvendigheden af en justering af det valgte forsøgsdesign for identifikation af markører blandt disse parametre.
Undersøgelserne generelt var baseret på uddragning af overensstemmelsen mellem to mulige udfald for samtlige variable i de valgte datasæt, hvilket er et særdeles restriktivt udfaldsrum. Selvom clustering som data mining metode er mindre følsom overfor støj, vil støj dog fortsat være en forstyrrende faktor. Valg af forskellige støjreducerende tiltag ved undersøgelsen af selektionssnit af journalnotater har dog ikke medvirket til at skabe klarhed over denne faktors betydning.
Slutteligt skal det dog pointeres, at yderligere undersøgelser vil være nødvendige for at uddrage betydningen af de valgte parametre og betydningen af informationer gemt i større databaser som EPJ-databasen for behandlingsudfald ved den antipsykotiske behandling. På baggrund af de gennemførte undersøgelser må jeg således konkludere, at det ikke er muligt at identificere markører blandt de valgte parametre som prædikterende for behandlingsudfaldene udskrivning og behandlingsskifte.

9.1 Perspektivering

Videnuddragning fra massen af journalnotater, som er registreret i EPJ-systemet på Psykiatrisk Center Sct. Hans er ikke nogen triviel opgave. Til trods for at de færreste vil betvivle værdien af den viden, der potentielt set genereres til dagligt i landets patientadministrationssystemer, synes de nuværende initiativer indenfor udviklingen af elektroniske patientjournaler - inklusive det system, som indenfor få år skal overtage pladsen fra det nuværende EPJ-system på Psykiatrisk Center Sct. Hans og regionens øvrige sygehuse og centre - at negligere betydningen af dette videnskatalog. Selvom dokumentationskravet i de elektroniske patientjournaler er opfyldt, er det langt fra nok til at danne grundlag for forskning i lidelser, i behandlingsregimer eller i de farmakologiske interventioner, som praktiseres på tværs af centrenes og sygehusenes sengeafsnit landet over.
Selvom en lang række metoder er blevet udviklet til at uddrage strukturer og relationer mellem tilnærmelsesvist uendelige datamængder uafhængig af det pågældende domæne eller datakilden, er der en væsentlig problemstilling knyttet til videnuddragning fra særligt patientadministrationssystemer: nemlig at kompleksiteten i anvendelsen af teknologien vil være en væsentlig hindring for dens anvendelse af klinikere

eller andre sundhedsfaglige eksperter.
Et væsentligt skridt i retning mod en løsning af problemet er at medtænke det forskningsmæssige perspektiv ved udformningen af fremtidens elektroniske patientjournaler. Såkaldte *click*-bokse, der tillader klinikere at indikere eksempelvis patientens symptombillede, vil udover at lette registreringsprocedurer også medvirke til at *strukturere* og ensarte og strukturere data og dermed lette brugen af disse data i forskningsøjemed samt klinikkens monitorering af patientens tilstand og dermed styrke evidensen i det daglige arbejde.

Litteratur

[Aichhorn *et al.*, 2000] W. Aichhorn, M. Gasser, E. M. Weiss, C. Adlassnig, and J. Marksteiner. Gender differences in pharmacokinetics and side effects of second generation antipsychotic drugs. *American Journal of Psychiatric Rehabilitation*, 4(2):199–215, 2000. 6

[Alavijeh *et al.*, 2005] M. S. Alavijeh, M. Chishty, M. Z. Qaiser, and A. M. Palmer. Drug metabolism and pharmacokinetics, the blood-brain barrier, and central nervous system drug discovery. *Journal of the American Society for Experimental NeuroTherapeutics*, 2:554–571, 2005. 24

[Almenoff *et al.*, 2007] J. S. Almenoff, E. N. Pattishall, T. G. Gibbs, W. DuMouchel, S. J. Evans, and N. Yuen. Novel statistical tools for monitoring the safety of marketed drugs. *Clinical pharmacology and therapeutics*, 82(2):157–166, 2007. 2

[Apiquian *et al.*, 2004] R. Apiquian, A. Fresán, C. Fuente-Sandoval, R. Ulloa, and H. Nicolini. Survey on schizophrenia treatment in mexico: perception and antipsychotic prescription patterns. *British Medical Journal of psychiatry*, 4(12):1–7, 2004. 34

[Arana, 2000] G. W. Arana. An overview of side effects caused by typical antipsychotics. *Journal of clinical psychiatry*, 61:5–11, 2000. 32

[Awad and Voruganti, 2004] A. Awad and L. Voruganti. New antipsychotics, compliance, quality of life, and subjective tolerability are patients better off? *Canadian Journal of Psychiatry*, 49(5):297–302, 2004. 5

[Bate *et al.*, 2008] A. Bate, M. Lindquist, and I. R. Edwards. The application of knowledge discovery in databases to post-marketing drug safety: example of the who database. *Fundamental clinical pharmacology*, 22(2):127–140, 2008. 2

[Berry *et al.*, 1999] M. W. Berry, Z. Drmac, and E. R. Jessup. Matrices, vector spaces, and information retrieval. *Society for Industrial and Applied Mathematics*, 41(2):335362, 1999. 61, 62, 63, 66

[Birkeland, 2007] S. F. Birkeland. Paranoia. *Ugeskrift for Læger*, 169(42):3566–3570, 2007. 22

[Bloch, 2006] I. Bloch. Duality vs adjunction and general form for fuzzy mathematical morphology. *Lecture Notes in Computer Science*, 3849:354–361, 2006. 93

[Buckley , 1993] C. Buckley . The importance of proper weighting methods. *Workshop on Human Language Technology*, Document retrieval and text retrieval:349–352, 1993. 61, 65

[Carlberger *et al.*,] J. Carlberger, H. Dalianis, M. Hassel, and O. Knutsson. 63

[Choi and Park, 2007] N. K. Choi and B. J. Park. Adverse drug reaction surveillance system in korea. *Journal of preventive medicine and public health*, 40(4):278–284, 2007. 2

[Chouinard *et al.*, 1978] G. Chouinard, B. D. Jones, and L. Annable. Neuroleptic-induced supersensitivity psychosis. *American Journal of Psychiatry*, 135(11):1409–1410, 1978. 11

[Cobaugh *et al.*, 2007] D. J. Cobaugh, A. R. Erdman, L. L. Booze, E. J. Scharman, G. Christianson, A. S. Manoguerra, E. M. Caravati, P. A. Chyka, A. D. Woolf, L. S. Nelson, and W. G. Troutman. Atypical antipsychotic medication poisoning: An evidence-based consensus guideline for out-of-hospital management. *Clinical Toxicology*, 45(8):918–942, 2007. 27

[Correll *et al.*, 2003] C. U. Correll, A. K. Malhotra, S. Kaushik, M. McMeniman, and J. M. Kane. Early prediction of antipsychotic response in schizophrenia. *American Journal of Psychiatry*, 160:2063–2065, 2003. 11

[Curtis *et al.*, 2008] V. A. Curtis, K. Katsafouros, H. Möller, R. Medori, and E. Sacchetti. Long-acting risperidone improves negative symptoms in stable psychotic patients. *Journal of Psychopharmacology*, 22(3):254–261, 2008. 30, 34

[Dazzan *et al.*, 2005] P. Dazzan, K. D. Morgan, K. Orr, G. Hutchinson, X. Chitnis, J. Suckling, P. Fearon, P. K. McGuire, R. M. Mallett, P. B. Jones, J. Leff, and R. M. Murray. Different effects of typical and atypical antipsychotics on grey matter in first episode psychosis: the Æsop study. *Neuropsychopharmacology*, 30:765774, 2005. 27

[Deloitte, 2007] Deloitte. Strategiske udviklingsveje for epj. Bestyrelsen for den nationale epj-organisation. Lokaliseret den 18. november 2008 på http://www.regioner.dk/Aktuelt/Nyheder/Nyheder%202007/Nyheder%20april%202007/ / media/migration%20folder/upload/filer/epjorganisation/ strategiskeudviklingsvejeepj$_a frapportering$%20$april$2007$.pdf.ashx$, 2007. 42

[Durante *et al.*, 2007] F. Durante, E. P. Klement, R. Mesiar, and C. Sempi. Conjunctors and their residual implicators: Characterizations and construction methods. *Mediterranean Journal of Mathematics*, 4(3):343–356, 2007. 93

[Emsley *et al.*, 2006] R. Emsley, J. Rabinowitz, and R. Medori. Time course for antipsychotic treatment response in first-episode schizphrenia. *American Journal of Psychiatry*, 163:743–745, 2006. 11

[Faries *et al.*, 2005] D. Faries, H. Ascher-Svanum, B. Zhu, C. Correll, and J. Kane. Antipsychotic monotherapy and polypharmacy in the naturalistic treatment of schizophrenia with atypical antipsychotics. *British Medical Journal of psychiatry*, 5(26):1–11, 2005. 34, 36

[Fayyad *et al.*, 1996] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *American Association for Artificial Intelligence*, 17(3):37–54, 1996. 51, 52, 53, 56

[Fog, 1995] R. Fog. Sct. hans hospital - træk af det ældste danske psykiatriske hospitals historie. Bibliotek for Læger. Lokaliseret den 11. november 2008 på http://www.psykiatri-regionh.dk/NR/ rdonlyres/F31A7DEF-A1E2-43EE-AFE1-A8C2BABA506F/0/ SctHansHospitaltraekafdetaeldstedanskepsykiatriskehospitalshistorie.pdf, 1995. 40

[Fowler, 2000] D. Fowler. Cognitive behavior therapy for psychosis: From understanding to treatment. *American Journal of Psychiatric Rehabilitation*, 4(2):199–215, 2000. 4

[Frangou and Byrne, 2000] S. Frangou and P. Byrne. How to manage the first episode of schizophrenia - early diagnosis and treatment may prevent social disability later. *British Medical Journal*, 321(7260):522523, 2000. 20, 23

[Freudenreich and Goff, 2002] O. Freudenreich and D. C. Goff. Antipsychotic combination therapy in schizophrenia. a review of efficacy and risks of current combinations. *Acta Psychiatrica Scandinavica*, 106:323–330, 2002. 36, 37

[Gaebel *et al.*, 2007] W. Gaebel, M. Riesbeck, B. Janssen, F. Schneider, T. Held, H. Mecklenburg, and H. Sas. Atypical and typical neuroleptics in acute schizophrenia and related delusional disorders - drug choice, switching outcome under naturalistic treatment conditions. *European Archives of Psychiatry and Clinical Neuroscience*, 253:175–184, 2007. 24

[Glenthøj *et al.*, 2008] B. Glenthøj, L. Peacock, A. Fink-Jensen, H. Lublin, A. K. Pagsberg, and T. Warrer. Tværgående vejledning for behandling af patienter med skizofreni og psykotiske tilstande med antipsykotika. Region Hovedstadens Psykiatri, 2008. 35

[Greenfield *et al.*, 1996] M. L. Greenfield, J. E. Kuhn, and E. M. Wojtys. A statistics primer. *American Journal of Sports Medicine*, 24(3):1–3, 1996. 12

[Han and Kamber, 2001] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, CA, 2001. 53, 55, 56, 57, 58, 59

[Hill *et al.*, 1992] C. Hill, N. A. Keks, H. Jackson, J. Kulkarni, D. Hannah, D. Copolov, and B. Singh. Symptomatic response to antipsychotics differs between recent onset and recurrent chronic schizophrenic patients. *Australian and New Zealand Journal of Psychiatry*, 26:417–422, 1992. 11

[Honer *et al.*, 2006] W. G. Honer, A. E. Thornton, E. Y. Chen, R. C. Chan, J. O. Wong, A. Bergmann, P. Falkai, E. Pomarol-Clotet, P. J. McKenna, E. Stip, R. Williams, G. W. MacEwan, K. Wasan, and R. Procyshyn. Clozapine alone versus clozapine and risperidone with refractory schizophrenia. *New England Journal of Medicine*, 354(5):472–482, 2006. 37

[Iskander *et al.*, 2006] J. Iskander, V. Pool, W. Zhou, and R. English-Bullard. Data mining in the us using the vaccine adverse event reporting system. *Drug safety : an international journal of medical toxicology and drug experience*, 29(5):375–384, 2006. 2

[Jibson and Tandon, 1998] M. D. Jibson and R. Tandon. New atypical medications. *Journal of Psychiatric Research*, 32:215–228, 1998. 24, 34

[Josiassen *et al.*, 2005] R. C. Josiassen, A. Joseph, E. Kohegyi, S. Stokes, M. Dadvand, W. W. Paing, and R. A. Shaughnessy. Clozapine augmented with risperidone in the treatment of schizophrenia: A randomized, double-blind, placebo-controlled trial. *American Journal of Psychiatry*, 162:130–136, 2005. 36, 37

[Kannabiran and Singh, 2008] M. Kannabiran and V. Singh. Metabolic syndrome and atypical antipsychotics: A selective literature review. *German journal of psychiatry*, 11(3):111–122, 2008. 33

[Kjeldsen, 2002] S. B. Kjeldsen. Første hospital med fuld epj-dækning. Sygeplejersken, 49, 2002. 8

[Kowalski, 1997] G. J. Kowalski. *Information Retrieval Systems: Theory and Implementation*. Kluwer Academic Publishers, Norwell, MA, 1997. 66

[Kreyenbuhl *et al.*, 2007] J. A. Kreyenbuhl, M. Valenstein, J. F. McCarthy, D. Ganoczy, and F. C. Blow. Long-term antipsychotic polypharmacy in the va health system: Patient characteristics and treatment patterns. *Psychiatr Services*, 58:489–495, 2007. 35

[Lader, 1999] M. Lader. Some adverse effects of antipsychotics: prevention and treatment. *Journal of clinical psychiatry*, 60:18–21, 1999. 36

[Lambert and Castle, 2003] T. J. Lambert and D. J. Castle. Pharmacological approaches to the management of schizophrenia. *Medical Journal of Australia*, 178:57–61, 2003. 18, 19, 26, 27

[Leitão-Azevedo *et al.*, 2006] C. L. Leitão-Azevedo, L. R. Guimarães, M. G. Belmonte-de-Abreu, C. S. Gama, M. I. Lobato, and P. S. Belmonte-de-Abreu. Increased dyslipidemia in schizophrenic outpatients using new generation antipsychotics. *Brasileiro de Psiquiatria*, 28(4):301–304, 2006. 33

[Lewis and Lieberman, 2000] D. A. Lewis and J. A. Lieberman. Catching up on schizophrenia: Natural history and neurobiology. *Neuron*, 28:325334, 2000. 20, 21, 26

[Lieberman *et al.*, 2005] J. A. Lieberman, T. S. Stroup, J. P. McEvoy, M. S. Swartz, R. A. Rosenheck, D. O. Perkins, R. S. Keefe, S. M. Davis, C. E. Davis, B. D. Lebowitz, J. Severe, and J. K. Hsiao. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *New England Journal of Medicine*, 353(12):1209–1223, 2005. 26, 27

[Liu and Loh, 2007] Y. Liu and H. Loh. A simple probability based term weighting scheme for automated text classification. *Lecture Notes in Computer Science*, 4570:33–43, 2007. 66

[Love, 2002] R. Love. Strategies for increasing treatment compliance: The role of long-acting antipsychotics. *American Journal of Health-System Pharmacists*, 59:10–15, 2002. 35, 36, 37

[Manchanda and Hirsch, 1986] R. Manchanda and S. R. Hirsch. Does propranolol have an antipsychotic effect? - a placebo-controlled study in acute schizophrenia. *British Journal of Psychiatry*, 148:701–707, 1986. 11

[Martínez-Fernández *et al.*, 2004] J. L. Martínez-Fernández, A. García-Serrano, P. Martínez, and J. Villena. Automatic keyword extraction for news finder. *Lecture Notes in Computer Science*, 3094:99–119, 2004. 65

[Masand, 2005] P. S. Masand. A review of pharmacologic strategies for switching to atypical antipsychotics. *Primary Care Companion to The Journal of Clinical Psychiatry*, 7(3):121–129, 2005. 36

[Mauri *et al.*, 2005] M. C. Mauri, F. Regispani, S. Beraldo, L. S. Volonteri, V. M. Ferrari, A. Fiorentini, and G. Invernizzi. Patterns of clinical use of antipsychotics in hospitalized psychiatric patients. *Progress in neuro-psychopharmacology & biological psychiatry*, 29(6):957–963, 2005. 8, 17, 34

[McCreadie and MacDonald, 1977] R. G. McCreadie and I. M. MacDonald. High dosage haloperidol in chronic schizophrenia. *British Journal of Psychiatry*, 131:310–316, 1977. 27, 28, 29

[Meystre *et al.*, 2008] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, 17:128–144, 2008. 2

[Miller and Craig, 2002] A. L. Miller and C. S. Craig. Combination antipsychotics: Pros, cons, and questions. *Schizophrenia Bulletin*, 28(1):105–109, 2002. 11, 35, 36

[Möller *et al.*, 2008] H. Möller, M. Riedel, M. Jäger, F. Wickelmaier, W. Maier, K. Kühn, G. Buchkremer, I. Heuser, J. Klosterkötter, M. Gastpar, D. F. Braus, R. Schlösser, F. Schneider, C. Ohmann, M. Riesbeck, and W. Gaebel. Short-term treatment with risperidone or haloperidol in first-episode schizophrenia: 8-week results of a randomized controlled trial within the german research network on schizophrenia. *International Journal of Neuropsychopharmacology*, 11:985–997, 2008. 28, 30

[Morrato *et al.*, 2007] E. H. Morrato, S. Dodd, G. Oderda, D. G. Haxby, R. Allen, and R. J. Valuck. Prevalence, utilization patterns, and predictors of antipsychotic polypharmacy: Experience in a multistate medicaid population, 1998-2003. *Clinical Therapeutics*, 29(3):183–195, 2007. 35

[Morris and Maisto, 2000] C. G. Morris and A. A. Maisto. *Psychology: An Introduction.* Prentice-Hall, Inc., New Jersey, NJ, 2000. 26

[Mullins *et al.*, 2006] I. M. Mullins, M. S. Siadaty, J. Lyman, K. Scully, C. T. Garrett, W. G. Miller, R. Muller, B. Robson, C. Apte, W. Sholom, I. Rigoutsos, D. Platt, S. Cohen, and W. A. Knaus. Data mining and clinical data repositories: Insights from a 667,000 patient data set. *Computers in Biology and Medicine*, 36:1351–1377, 2006. 3

[Neergaard, 2005] L. D. Neergaard. Behandling af psykotiske patienter med misbrugsproblemer i h:s. H:S Sundhedsfagligt Råd for Psykiatri. Lokaliseret den 18. november 2008 på http://www.hosp.dk/direktion.nsf/pics/Dobbeltdiagnose%20endelig.pdf/$FILE/Dobbeltdiagnose$%20$endelig.pdf$, 2005. 41

[Nelson *et al.*, 2001] E. B. Nelson, E. Rielage, J. A. Welge, and P. E. Keck. An open trial of olanzapine in the treatment of patients with psychotic depression. *Annals of Clinical Psychiatry*, 13:147–151, 2001. 30

[Newcomer, 2004] J. W. Newcomer. Understanding the risk factors in antipsychotic treatment. *Advanced Studies in Medicine*, 4(10H):1059–1064, 2004. 6

[Nordland, 2001] O. Nordland. Den elektroniske patientjournal kan også blive borte. *Ugeskrift for Læger*, 163(19):2698, 2001. 1

[Okasha and Tewfik, 1964] A. Okasha and G. I. Tewfik. Haloperidol: A controlled clinical trial in chronic disturbed psychotic patients. *British Journal of Psychiatry*, 110:56–60, 1964. 27, 29

[Pantelis and Lambert, 2003] C. Pantelis and T. J. Lambert. Managing patients with „treatment-resistant" schizophrenia. *Medical Journal of Australia*, 178:62–666, 2003. 19

[Patel and David, 2005] M. X. Patel and A. S. David. Why arent depot antipsychotics prescribed more often and what can be done about it? *Advances in Psychiatric Treatment*, 11:203–213, 2005. 25

[Pedersen *et al.*, 2008] C. Pedersen, L. Bjerrum, K. P. Dalhoff, H. Friis, and J. Hendel. Antipsykotika. Lokaliseret den 26. november 2008 på http://medicin.dk/(p25dpu45obdkkfucbu4zwzj4)/show.aspx, 2008. 10, 25, 28, 31, 71

[Peuskens, 1995] J. Peuskens. Risperidone in the treatment of patients with chronic schizophrenia: a multi-national, multi-centre, double-blind, parallel-group study versus haloperidol. *British Journal of Psychiatry*, 166:712–726, 1995. 28, 31

[Prather et al., 1997] H. Prather et al. Medical data mining: Knowledge discovery in a clinical data warehouse. *AMIA Annual Fall Symposium*, 445:101–5, 1997. 50

[Razvan and Bunescu, 2005] R. M. Razvan and M. Bunescu. Mining knowledge from text using information extraction. *Special issue on Text Mining and Natural Language Processing*, 7(1):3–10, 2005. 50

[Region H, 2007] Region H. Strukturplan for ledelsen i region hovedstadens virksomheder. Regionsrådet. Lokaliseret den 14. november 2008 på http://www.regionh.dk/NR/rdonlyres/ 04EC8EBE-C906-470C-9B27-906EC8FA8A91/0/StrukturplanforledelseniRegionHovedstadensvirksomheder.pdf, 2007. 39

[Remvig and Sonne, 1961] J. Remvig and L. M. Sonne. Chlorprothixene („truxal") compared to chlorpromazine. *Psychopharmacologia*, 2:203–208, 1961. 27, 28, 29

[Riedel *et al.*, 2007] M. Riedel, N. Müller, I. Spellmann, R. R. Engel, R. Musil, R. Valdevit, S. Dehning, A. Douhet, A. Cerovecki, M. Strassnig, and H. Möller. Efficacy of olanzapine versus quetiapine on cognitive dysfunctions in patients with an acute episode of schizophrenia. *European Archives of Psychiatry and Clinical Neuroscience*, 257(7):402–412, 2007. 30, 34

[Romesburg, 2004] C. Romesburg. *Cluster Analysis for Researchers*. Lulu.com, Morrisville, NC, 2004. 67

[Rubio *et al.*, 2006a] G. Rubio, I. Martinez, G. Ponce, M. A. Jimenez-Arriero, and F. Lopez-Munoz. Long-acting injectable risperidone compared with zuclopenthixol in the treatment of schizophrenia with substance abuse comoridity. *Canadian Journal of Psychiatry*, 51(8):531–539, 2006. 28, 31, 34

[Rubio *et al.*, 2006b] G. Rubio, I. Martinez, A. Recio, G. Ponce, F. Lopez-Munoz, C. Alamo, M. A. Jimenez-Arriero, and T. Palomo. Long-acting risperidone improves negative symptoms in stable psychotic patients. *European Journal of Psychiatry*, 20(3):133–146, 2006. 30

[Salton, 1968] G. Salton. *Automatic information organization and retrieval.* McGraw Hill, New York, NY, 1968. 61

[Sanger *et al.*, 1999] T. M. Sanger, J. A. Lieberman, M. Tohen, S. Grundy, C. Beasley, and G. D. Tollefson. Olanzapine versus haloperidol treatment in first-episode psychosis. *American Journal of Psychiatry*, 156:79–87, 1999. 20, 23, 28, 30, 34

[Schneider, 2007] S. Schneider. Sct. hans nyt. Personaleblad for Sct. Hans Hospital, 2007. 39

[Schneider, 2008a] S. Schneider. Hs patientvisitation. Lokaliseret den 11. november 2008 på http://https://www.phsinfo.dk/__41256a850068d90d.nsf/82e3b8a9dcaf353641256a88002c69b7!OpenViewStart=1Count=3000 Expand=88, 2008. 39

[Schneider, 2008b] S. Schneider. Information om psykiatrisk center sct. hans hos region hovedstadens psykiatri. Lokaliseret den 11. november 2008 på http://http://www.psykiatri-regionh.dk/menu/Centre/Psykiatriske+centre/Psykiatrisk+Center+ Sct.+Hans/, 2008. 39, 40

[Schooler *et al.*, 2005] N. Schooler, J. Rabinowitz, M. Davidson, R. Emsley, P. D. Harvey, L. Kopala, P. D. McGorry, I. V. Hove, M. Eerdekens, W. Swyzen, and G. D. Smedt. Risperidone and haloperidol in first-episode psychosis: A long-term randomized trial. *American Journal of Psychiatry*, 162:947–953, 2005. 28, 31

[Sebbelov, 2001] K. B. Sebbelov. Elektronisk patientjournal på sct. hans. Sygeplejersken, 43, 2001. 41

[Sekine *et al.*, 1999] Y. Sekine, T. Rikihisa, H. Ogata, H. Echizen, and Y. Arakawa. Correlations between in vitro affinity of antipsychotics to various neurotransmitter receptors and clinical incidence of their adverse drug reactions. *European Journal of Clinical Pharmacology*, 55:583–587, 1999. 24

[Serretti *et al.*, 2004] A. Serretti, D. Ronchi, C. Lorenzi, and D. Berardi. New antipsychotics and schizophrenia: A review on efficacy and side effects. *Current medical chemistry*, 11:343–358, 2004. 32, 33

[Silberschatz *et al.*, 2006] A. Silberschatz, H. F. Korth, and S. Sudarshan. *Database System Concepts, 5th edition.* McGraw Hill, New York, NY, 2006. 45

[Smith, 2006] E. Smith. Bekendtgørelse om lægers, tandlægers, kiropraktorers, jordemødres, kliniske diætisters, kliniske tandteknikeres, tandplejeres, optikeres og kontaktlinseoptikeres patientjournaler (journalføring, opbevaring, videregivelse og overdragelse m.v.). Lokaliseret den 1. juni 2008 på https://www.retsinformation.dk/Forms/R0710.aspx?id=11055exp=1, 2006. 1, 41

[Stefansson *et al.*, 2008] H. Stefansson, D. Rujescu, S. Cichon, O. Pietiläinen, A. Ingason, S. Steinberg, R. Fossdal, E. Sigurdsson, T. Sigmundsson, J. E. Buizer-Voskamp, T. Hansen, K. D. Jakobsen, P. Muglia1, C. Francks, P. M. Matthews, C. A. Gylfason, B. V. Halldorsson, D. Gudbjartsson, T. E. Thorgeirsson, A. Sigurdsson, A. Jonasdottir, A. Bjornsson, S. Mattiasdottir, T. Blondal, M. Haraldsson, B. B. Magnusdottir, I. Giegling, H. Möller, A. Hartmann, K. V. Shianna, D. Ge, A. C. Need, C. Crombie, G. Fraser, N. Walker, J. Lonnqvist, J. Suvisaari, A. Tuulio-Henriksson, T. Paunio, T. Toulopoulou, E. Bramon, M. D. Forti, R. Murray, M. Ruggeri, E. Vassos, S. Tosato, D. Rujescu, D. Rujescu, D. Rujescu, D. Rujescu, D. Rujescu, D. Rujescu, M. Walshe, T . Li, C. Vasilescu, T. W. Mühleisen, A. G. Wang, H. Ullum, S. Djurovic, I. Melle, J. Olesen, L. A. Kiemeney, B. Franke, C. Sabatti, N. B. Freimer, J. R. Freimer, U. Thorsteinsdottir, A. Kong, O. A. Andreassen, R. A. Ophoff, A. Georgi, M. Rietschel, T. Werge, H. Petursson, D. B. Goldstein, M. M. Nöthen, L. Peltonen, D. A. Collier, D. S. Clair, and K. Stefansson. Large recurrent microdeletions associated with schizophrenia. *Nature*, 445:232–236, 2008. 21

[Stenstrøm *et al.*, 2008] A. D. Stenstrøm, B. Dehlholm-Lambertsen, and Nøhr-Jensen P. Tidlige skizofreniforme symptomer hos børn. *Ugeskrift for Læger*, 170(15):1227–1232, 2008. 21

[Tranulis *et al.*, 2008] C. Tranulis, L. Skalli, P. Lalonde, L. Nicole, and E. Stip. Benefits and risks of antipsychotic polypharmacy - an evidence-based review of the literature. *Drug Safety*, 31(1):7–20, 2008. 11

[Trifiro *et al.*, 2005] G. Trifiro, E. Spina, O. Brignoli, E. Sessa, A. P. Caputi, and G. Mazzaglia. Antipsychotic prescribing pattern among italian general practitioners: a population-based study during the years 19992002. *European Journal of Clinical Pharmacololgy*, 61:4753, 2005. 34

[Tsai and Bond, 2008] J. Tsai and G. Bond. A comparison of electronic records to paper records in mental health centers. *International Journal for Quality in Health Care*, 20(2):136–143, 2008. 1

[Viera and Garrett, 2005] A. J. Viera and J. M. Garrett. Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37(5):360–363, 2005. 94, 96

[Vingtoft et al., 2005] S. Vingtoft et al. Epj-observatoriet statusrapport 2005. EPJ-Observatoriet. Lokaliseret den 30. september 2008 på http://www.sst.dk/upload/informatik_og_sundhedsdata/sundhedsinformatik /epj/-statusrapport2005.pdf, 2005. 1

[WHO, 1993] WHO. The icd-10 classification of mental and behavioural disorders. diagnostic criteria for research. Geneva. WHO, 1993. 8, 17, 18, 21, 22, 23

Bilag A

Dataindsamling

Datafremstillinger i projektet er fremkommet ved brug af SQL-forespørgsler i EPJ-systemets database på Psykiatrisk Center Sanct Hans. Forespørgslerne er formuleret til enten systemets eksisterende struktur eller afledte tabeller. For nærmere information om baggrunden for oprettelsen af specifikke afledte tabeller henvises til det pågældende afsnit i rapporten. SQL-forespørgslerne er anvendt enten direkte på databasens tabeller via et DB2-kommandovindue-input, på de afledte tabeller via Oracle SQL Developer eller indirekte via javainstruktioner de steder hvor outputgenereringen har været betinget af en mere avanceret analyse.

A.1 Deskriptive datafremstillinger for hele patientpopulationen

A.1.1 Antal journalnotater i 2007

SQL-forespørgsel:

```
SELECT count(distinct nh.noteid) FROM noteheader nh,
notes no WHERE nh.noteid=no.noteid AND to_date(
nh.notetime, 'YYYY-MM-DD') >= to_date('2007-01-01',
'YYYY-MM-DD') AND to_date(nh.notetime, 'YYYY-MM-DD')
< to_date('2008-01-01', 'YYYY-MM-DD');
```

Output: 351064

A.1.2 Antal tilknyttede patienter i 2007

SQL-forespørgsel:

```
SELECT count(DISTINCT pr.pid) FROM prescription pr, ipe
WHERE to_date(ipe.registrationdt, 'YYYY-MM-DD') <
to_date('2008-01-01', 'YYYY-MM-DD') AND (to_date(
ipe.dischargedt, 'YYYY-MM-DD') >= to_date('2007-01-01',
```

```
'YYYY-MM-DD') OR ipe.dischargedt IS NULL) AND pr.pid =
ipe.pid AND pr.status <> 9 AND to_date(pr.prescstart,
'YYYY-MM-DD') >= to_date('2007-01-01', 'YYYY-MM-DD') AND
to_date(pr.prescstart, 'YYYY-MM-DD') < to_date(
'2008-01-01', 'YYYY-MM-DD');
```

Output: 783

A.1.3 Registrerede patienter i EPJ-systemet fra 1. april 2003 frem til 1. april 2008

SQL-forespørgsel:

```
SELECT COUNT (DISTINCT pe.pid) patients FROM person pe,
prescription pr, ipe WHERE pe.pid = pr.pid
AND pe.pid=ipe.pid AND pr.prescstart <= '2008-04-01'
AND ipe.registrationdt <= '2008-04-01'
AND pr.prescstop >= '2003-04-01' AND ipe.dischargedt
<= '2008-04-01'AND pr.status<>9 AND pr.prescord<>5;
```

Output: 1987

A.1.4 Patienter registreret i EPJ-systemet fra 1. april 2003 frem til 1. april 2008 fordelt på køn

SQL-forespørgsel:

```
SELECT COUNT(DISTINCT pe.pid) PATIENTS, pe.SEX FROM
person pe, prescription pr, ipe WHERE
pe.pid = pr.pid AND pe.pid=ipe.pid AND pr.prescstart
<= '2008-04-01' AND ipe.registrationdt <= '2008-04-01'
AND pr.prescstop >= '2003-04-01' AND ipe.dischargedt
>= '2003-04-01'AND pr.status<>9 AND pr.prescord<>5
GROUP BY pe.sex;
```

Output: Tabel A.1

Antal	Køn
667	Kvinder
1320	Mænd

Tabel A.1: Patienter fordelt på køn

ICD-10	Antal kvinder	Antal mænd
F 10-19	80	200
F 20-29	96	265
F 30-39	49	61

Tabel A.2: F1-F3 diagnosespektrets fordeling for de identificerede patienter

A.2 Deskriptive datafremstillinger for det empiriske grundlag

A.2.1 Antal identificerede patienter med tilfælde af 4 ugers antipsykotisk monoterapi

SQL-forespørgsel:

```
SELECT COUNT(DISTINCT pid) FROM ipe, ipes
WHERE ipe.ipeid=ipes.ipeid;
```

Output: 697

A.2.2 F1-F3 diagnosespektrets fordeling for de identificerede patienter

SQL-forespørgsel:

```
SELECT COUNT(pid) patienter, sex, diagnoser FROM (
SELECT DISTINCT dp.pid, pe.sex, SUBSTR(dp.code, 1, 2)
diagnoser FROM person pe, diagnosis dp, ipe, ipes,
drugtregimen dtr WHERE dp.pid = pe.pid AND ipe.pid =
pe.pid AND ipe.ipeid = ipes.ipeid AND dtr.groupid =
ipes.groupid)GROUP BY sex, diagnoser ORDER BY patienter
DESC;
```

Output: Tabel A.2

A.3 Deskriptive datafremstillinger for de valgte behandlingsforløb

A.3.1 Antal patienter indeholdt i behandlingsskifte-gruppen

SQL-forespørgsel:

```
SELECT COUNT(*) FROM mono_skift;
```

Output: 130

A.3.2 Antal patienter indeholdt i udskrivnings-gruppen

SQL-forespørgsel:

```
SELECT COUNT(*) FROM mono_ud;
```

Output: 224

A.3.3 Behandlingsvarighed i dage for patientkohortens grupper

SQL-forespørgsel:

```
SELECT distinct dtr.treat_id, end-begin varighed, cli.name
FROM clinic cli, ipe, ipes, drugtregimen dtr, mono_skift ms
WHERE dtr.groupid = ipes.groupid AND ipe.ipeid =
ipes.ipeid AND ipe.clinicid = cli.clinicid
AND begin < to_date(SUBSTR(ipe.dischargedt, 1, 10),
'yyyy-mm-dd') AND end > to_date(SUBSTR(
ipe.registrationdt, 1, 10), 'yyyy-mm-dd')
AND ms.treat_id = dtr.treat_id;
```

A.3.4 Antal behandlingforløb ved den samlede indlæggelse for hver af patientkohortens grupper

SQL-forespørgsel:

```
select dtr.groupid, count(treat_id) antal, cli.name name
FROM drugtregimen dtr, clinic cli, ipe, ipes WHERE
dtr.groupid IN (SELECT groupid FROM drugtregimen dtr,
mono_skift mu WHERE dtr.treat_id = mu.treat_id)
AND dtr.groupid = ipes.groupid AND ipe.ipeid =
ipes.ipeid AND ipe.clinicid = cli.clinicid
AND begin < to_date(SUBSTR(ipe.dischargedt, 1, 10),
'yyyy-mm-dd') AND end > to_date(SUBSTR(
ipe.registrationdt, 1, 10), 'yyyy-mm-dd')
GROUP BY dtr.groupid, cli.name;
```

A.3.5 Kønsfordeling i patientkohortens grupper

SQL-forespørgsel:

```
select distinct dtr.treat_id, sex, cli.name name FROM
clinic cli, person pe, ipe, ipes, drugtregimen dtr,
mono_ud mu, diagnosis dg WHERE mu.treat_id = dtr.treat_id
AND dtr.groupid = ipes.groupid AND ipes.ipeid = ipe.ipeid
AND pe.pid = ipe.pid AND ipe.clinicid = cli.clinicid;
```

Bilag B

Dataanalyser

B.1 Behandlingsforløb

B.1.1 Patienter i monoterapi

Behandlingsskift: Tabellen mono_skift

```
CREATE TABLE mono_skift AS SELECT DISTINCT t.treat_id
FROM treatments t, drugtregimen dtr, (SELECT groupid,
MIN(treat_id) treat_id FROM treatments GROUP BY groupid
) s, (SELECT groupid, COUNT(treat_id) antal_beh FROM
treatments GROUP BY groupid) b,
(SELECT MIN(groupid) groupid, pid FROM ipes, ipe WHERE
ipes.ipeid = ipe.ipeid GROUP BY pid) c WHERE
s.treat_id = t.treat_id AND antal_spor=1 AND
s.groupid=b.groupid AND c.groupid=b.groupid AND
antal_beh >1 AND s.treat_id = dtr.treat_id;
```

```
CREATE INDEX mono_skift_treat_id
ON mono_skift (treat_id);
```

Udskrivninger: Tabellen mono_ud

```
CREATE TABLE mono_ud AS SELECT distinct t.treat_id
FROM treatments t, adms, drugtregimen dtr,
(SELECT groupid, MAX(treat_id) treat_id FROM treatments
GROUP BY groupid) s, (SELECT MIN(groupid) groupid, pid
FROM ipes, ipe WHERE ipes.ipeid = ipe.ipeid
GROUP BY pid) c WHERE s.treat_id = t.treat_id AND
antal_spor=1 AND s.groupid=adms.groupid AND
s.groupid=c.groupid AND t.treat_id NOT IN
(SELECT distinct treat_id FROM ipe, ipes,
drugtregimen dtr WHERE dtr.groupid = ipes.groupid AND
```

```
ipes.ipeid = ipe.ipeid AND pid in (SELECT pid FROM ipe,
ipes, drugtregimen dtr WHERE ipe.ipeid = ipes.ipeid AND
ipes.groupid = dtr.groupid AND dtr.treat_id in
(SELECT treat_id FROM mono_skift))) AND
s.treat_id = dtr.treat_id AND enddt-end<=9;
```

```
CREATE INDEX mono_ud_treat_id
ON mono_ud (treat_id);
```

B.2 Idf-vægtning af journalnotater

B.2.1 Uden lemmatisering

Det fulde dokumentsæt

Ordtabel: Tabellen doc_words

```
CREATE TABLE doc_words AS SELECT DISTINCT dr.treat_id,
aw.word, aw.noteid, aw.fieldid, aw.pos FROM
DRUGTREATMENTSETS ds, DRUGTREGIMEN dr, ALL_WDOCS aw,
notes no, noteheader nh, prescription pr,
DRUGSPRESCRIBED dp WHERE to_date(nh.notetime,
'YYYY-MM-DD') >= dr.begin AND to_date(nh.notetime,
'YYYY-MM-DD') <= dr.end AND no.noteid=nh.noteid AND
no.pid = pr.pid AND dr.treat_id=ds.treat_id AND
ds.treat_id IN (SELECT treat_id FROM mono_ud mu UNION
SELECT treat_id FROM mono_skift ms) AND ds.DT_ID=
dp.DT_ID AND dp.PRESCRIPTIONID=pr.PRESCRIPTIONID AND
no.noteid=aw.noteid;
```

```
CREATE INDEX doc_words_treat_id ON doc_words(treat_id);
CREATE INDEX doc_words_word ON doc_words(word);
```

Termfrekvens: Tabellen mono_tf

```
CREATE TABLE mono_tf AS
(SELECT treat_id, word, COUNT(*) tf FROM doc_words
GROUP BY treat_id, word);
```

```
CREATE INDEX mono_tf_treat_id ON mono_tf(treat_id);
CREATE INDEX mono_tf_word ON mono_tf(word);
```

Termnormalisering: Tabellen mono_max_tf

```
CREATE TABLE mono_max_tf AS (SELECT treat_id, MAX(tf)
maxtf FROM (SELECT * FROM mono_tf) GROUP BY treat_id);
```

```
CREATE INDEX mono_max_tf_treat_id ON
mono_max_tf(treat_id);
```

Dokumentfrekvens: Tabellen mono_df

```
CREATE TABLE mono_df AS (SELECT word, COUNT(*) df FROM
(SELECT DISTINCT treat_id, word FROM doc_words)
GROUP BY word);
```

```
CREATE INDEX mono_df_word ON mono_df(word);
```

Elementoptælling: Tabellen mono_n

```
CREATE TABLE mono_n AS
(SELECT COUNT(DISTINCT treat_id) n FROM doc_words);
```

Inverteret dokumentfrekvens: Tabellen mono_wind

```
CREATE TABLE mono_wind AS
SELECT mono_tf.treat_id, mono_tf.word, (tf/maxtf)*
(LOG(2, n/df)) weight FROM mono_tf, mono_max_tf,
mono_df, mono_n
WHERE mono_tf.treat_id=mono_max_tf.treat_id AND
mono_tf.word=mono_df.word;
```

```
CREATE INDEX mono_wind_treat_id ON mono_WIND(treat_id);
CREATE INDEX mono_wind_word ON mono_WIND(word);
```

Vektormål: Tabellen mono_dist

```
CREATE TABLE mono_DIST AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance(a.treat_id, b.treat_id),0) dist FROM
(SELECT DISTINCT treat_id FROM mono_WIND)a,
(SELECT DISTINCT treat_id FROM mono_WIND)b
WHERE a.treat_id < b.treat_id;
```

Afstandsberegning: Funktionen distance()

```
create or replace FUNCTION Distance(treat_id1 INTEGER,
treat_id2 INTEGER) RETURN NUMBER IS

result NUMBER := 0;

BEGIN
WITH ip AS
(SELECT SUM(weight) weight FROM
(SELECT word, SUM(weight) weight FROM
  (SELECT a.word, a.weight* b.weight weight FROM
   (SELECT word, weight FROM mono_WIND WHERE
   treat_id = treat_id1) a,
   (SELECT word, weight FROM mono_WIND WHERE
   treat_id = treat_id2) b
```

```
  WHERE a.word=b.word)
GROUP BY word)),
LENGTH AS
(SELECT SQRT(a.len * b.len) len FROM
(SELECT SUM(POWER(weight,2)) len FROM mono_WIND WHERE
treat_id = treat_id1) a,
(SELECT SUM(POWER(weight,2)) len FROM mono_WIND WHERE
treat_id = treat_id2) b)
SELECT weight/len INTO result FROM ip, LENGTH;

RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

Kun de sidste 10 notater

Ordtabel: Tabellen doc_words_last10

```
CREATE TABLE doc_words_last10 AS SELECT DISTINCT
t.treat_id, word, t.noteid, fieldid, pos FROM
(SELECT DISTINCT treat_id, noteid, RANK()
OVER(PARTITION BY treat_id ORDER BY noteid DESC) rn
FROM (SELECT DISTINCT treat_id, noteid FROM doc_words
WHERE treat_id IN (SELECT treat_id FROM mono_ud mu
UNION
SELECT treat_id FROM mono_skift ms) ORDER BY treat_id,
noteid DESC) ORDER BY rn ) t, doc_words dw WHERE rn
IN (1,2,3,4,5,6,7,8,9,10) AND t.noteid=dw.noteid;
```

```
CREATE INDEX doc_words_last10_treat_id ON
doc_words_last10(treat_id);
CREATE INDEX doc_words_last10_word ON
doc_words_last10(word);
```

Termfrekvens: Tabellen mono_last10_tf

```
CREATE TABLE mono_last10_tf AS
(SELECT treat_id, word, COUNT(*) tf
FROM doc_words_last10 GROUP BY treat_id, word);
```

```
CREATE INDEX mono_last10_tf_treat_id ON
mono_last10_tf(treat_id);
CREATE INDEX mono_last10_tf_word ON
```

```
mono_last10_tf(word);
```

Termnormalisering: Tabellen mono_last10_max_tf

```
CREATE TABLE mono_last10_max_tf AS
(SELECT treat_id, MAX(tf) maxtf FROM
(SELECT * FROM mono_last10_tf)
GROUP BY treat_id);
```

```
CREATE INDEX mono_last10_max_tf_treat_id ON
mono_last10_max_tf(treat_id);
```

Dokumentfrekvens: Tabellen mono_last10_df

```
CREATE TABLE mono_last10_df AS
(SELECT word, COUNT(*) df FROM
(SELECT DISTINCT treat_id, word FROM doc_words_last10)
GROUP BY word);
```

```
CREATE INDEX mono_last10_df_word ON
mono_last10_df(word);
```

Elementoptælling: Tabellen mono_last10_n

```
CREATE TABLE mono_last10_n AS
(SELECT COUNT(DISTINCT treat_id) n FROM
doc_words_last10);
```

Inverteret dokumentfrekvens: Tabellen mono_last10_wind

```
CREATE TABLE mono_last10_WIND AS
SELECT mono_last10_tf.treat_id, mono_last10_tf.word,
(tf/maxtf)*(LOG(2, n/df)) weight FROM mono_last10_tf,
mono_last10_max_tf, mono_last10_df, mono_last10_n
WHERE mono_last10_tf.treat_id=
mono_last10_max_tf.treat_id
AND mono_last10_tf.word=mono_last10_df.word;
```

```
CREATE INDEX mono_last10_wind_treat_id ON
mono_last10_WIND(treat_id);
CREATE INDEX mono_last10_wind_word ON
mono_last10_WIND(word);
```

Vektormål: Tabellen mono_last10_dist

```
CREATE TABLE mono_last10_DIST AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance_last10(a.treat_id, b.treat_id),0) dist
FROM (SELECT DISTINCT treat_id FROM mono_last10_WIND)a,
(SELECT DISTINCT treat_id FROM mono_last10_WIND)b
WHERE a.treat_id < b.treat_id;
```

Afstandsberegning: Funktionen distance_last10()

```
create or replace FUNCTION Distance_last10
(treat_id1 INTEGER, treat_id2 INTEGER) RETURN NUMBER IS

result NUMBER := 0;

BEGIN

WITH ip AS
(SELECT SUM(weight) weight FROM
(SELECT word, SUM(weight) weight FROM
  (SELECT a.word, a.weight* b.weight weight FROM
   (SELECT word, weight FROM mono_last10_WIND WHERE
   treat_id = treat_id1) a,
   (SELECT word, weight FROM mono_last10_WIND WHERE
   treat_id = treat_id2) b
  WHERE a.word=b.word)
GROUP BY word)),
LENGTH AS
(SELECT SQRT(a.len * b.len) len FROM
(SELECT SUM(POWER(weight,2)) len FROM mono_last10_WIND
        WHERE treat_id = treat_id1) a,
(SELECT SUM(POWER(weight,2)) len FROM mono_last10_WIND
        WHERE treat_id = treat_id2) b)
SELECT weight/len INTO result FROM ip, LENGTH;


RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

Kun de sidste 10% af samtlige notater

Klassisk IR undersøgelse med normalisering ved max frekvens og idf-vægtning

Ordtabel: Tabellen doc_words_last10perc

```
CREATE TABLE doc_words_last10perc AS SELECT DISTINCT
t.treat_id, word, t.noteid, fieldid, pos FROM (SELECT
DISTINCT treat_id, noteid, RANK() OVER(PARTITION BY
treat_id ORDER BY noteid DESC) rn FROM (
SELECT DISTINCT treat_id, noteid FROM doc_words WHERE
treat_id IN (SELECT treat_id FROM mono_ud mu UNION
SELECT treat_id FROM mono_skift ms) ORDER BY treat_id,
noteid DESC) ORDER BY rn) t, (SELECT treat_id,
count(DISTINCT noteid) n FROM doc_words GROUP BY
treat_id) tn, doc_words dw WHERE rn <=ROUND(0.1*n)
AND t.noteid=dw.noteid AND t.treat_id = tn.treat_id;
```

```
CREATE INDEX doc_words_last10perc_treat_id ON
doc_words_last10perc(treat_id);
CREATE INDEX doc_words_last10perc_word ON
doc_words_last10perc(word);
```

Termfrekvens: Tabellen mono_last10perc_tf

```
CREATE TABLE mono_last10perc_tf AS
(SELECT treat_id, word, COUNT(*) tf FROM
doc_words_last10perc
GROUP BY treat_id, word);
```

```
CREATE INDEX mono_last10perc_tf_treat_id ON
mono_last10perc_tf(treat_id);
CREATE INDEX mono_last10perc_tf_word ON
mono_last10perc_tf(word);
```

Termnormalisering: Tabellen mono_last10perc_max_tf

```
CREATE TABLE mono_last10perc_max_tf AS
(SELECT treat_id, MAX(tf) maxtf FROM
(SELECT * FROM mono_last10perc_tf)
GROUP BY treat_id);
```

```
CREATE INDEX mono_last10pc_max_tf_treat_id ON
mono_last10perc_max_tf(treat_id);
```

Dokumentfrekvens: Tabellen mono_last10perc_df

```
CREATE TABLE mono_last10perc_df AS (SELECT word,
COUNT(*) df FROM (SELECT DISTINCT treat_id, word
FROM doc_words_last10perc) GROUP BY word);
```

```
CREATE INDEX mono_last10perc_df_word
ON mono_last10perc_df(word);
```

Elementoptælling: Tabellen mono_last10perc_n

```
CREATE TABLE mono_last10perc_n AS
(SELECT COUNT(DISTINCT treat_id) n
FROM doc_words_last10perc);
```

Inverteret dokumentfrekvens: Tabellen mono_last10perc_wind

```
CREATE TABLE mono_last10perc_WIND AS
SELECT mono_last10perc_tf.treat_id,
mono_last10perc_tf.word, (tf/maxtf)*(LOG(2, n/df))
weight FROM mono_last10perc_tf, mono_last10perc_max_tf,
mono_last10perc_df, mono_last10perc_n WHERE
mono_last10perc_tf.treat_id=
mono_last10perc_max_tf.treat_id AND
mono_last10perc_tf.word=mono_last10perc_df.word;
```

```
CREATE INDEX mono_last10perc_wind_treat_id ON
mono_last10perc_WIND(treat_id);
CREATE INDEX mono_last10perc_wind_word ON
mono_last10perc_WIND(word);
```

Vektormål: Tabellen mono_last10perc_dist

```
CREATE TABLE mono_last10perc_DIST AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance_last10perc(a.treat_id, b.treat_id),0) dist
FROM
(SELECT DISTINCT treat_id FROM mono_last10perc_WIND)a,
(SELECT DISTINCT treat_id FROM mono_last10perc_WIND)b
WHERE a.treat_id < b.treat_id;
```

Afstandsberegning: Funktionen distance_last10perc()

```
create or replace FUNCTION
Distance_last10perc(treat_id1 INTEGER,
treat_id2 INTEGER) RETURN NUMBER IS

result NUMBER := 0;

BEGIN

WITH ip AS
(SELECT SUM(weight) weight FROM
(SELECT word, SUM(weight) weight FROM
  (SELECT a.word, a.weight* b.weight weight FROM
   (SELECT word, weight FROM mono_last10perc_WIND
   WHERE treat_id = treat_id1) a,
   (SELECT word, weight FROM mono_last10perc_WIND
   WHERE treat_id = treat_id2) b
  WHERE a.word=b.word)
GROUP BY word)),
LENGTH AS

(SELECT SQRT(a.len * b.len) len FROM
(SELECT SUM(POWER(weight,2)) len FROM
 mono_last10perc_WIND WHERE treat_id = treat_id1) a,
(SELECT SUM(POWER(weight,2)) len FROM
 mono_last10perc_WIND WHERE treat_id = treat_id2) b)
SELECT weight/len INTO result FROM ip, LENGTH;


RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

Kun lægenotater

Ordtabel: Tabellen doc_words_med

```
CREATE TABLE doc_words_med AS
SELECT * FROM doc_words WHERENoteid in
(SELECT DISTINCT noteid FROM noteheader nh WHERE
nh.noteheader like '%æge%');
```

```
CREATE INDEX doc_words_med_treat_id ON
doc_words_med(treat_id);
CREATE INDEX doc_words_med_word ON
doc_words_med(word);
```

Termfrekvens: Tabellen mono_med_tf

```
CREATE TABLE mono_med_tf AS
(SELECT treat_id, word, COUNT(*) tf FROM doc_words_med
GROUP BY treat_id, word);
```

```
CREATE INDEX mono_med_tf_treat_id ON
mono_med_tf(treat_id);
CREATE INDEX mono_med_tf_word ON mono_med_tf(word);
```

Termnormalisering: Tabellen mono_max_med_tf

```
CREATE TABLE mono_max_med_tf AS
(SELECT treat_id, MAX(tf) maxtf FROM (SELECT * FROM
mono_med_tf) GROUP BY treat_id);
```

```
CREATE INDEX mono_max_med_tf_treat_id ON
mono_max_med_tf(treat_id);
```

Dokumentfrekvens: Tabellen mono_med_df

```
CREATE TABLE mono_med_df AS (SELECT word, COUNT(*) df
FROM (SELECT DISTINCT treat_id, word FROM
doc_words_med) GROUP BY word);
```

```
CREATE INDEX mono_med_df_word ON mono_med_df(word);
```

Elementoptælling: Tabellen mono_med_n

```
CREATE TABLE mono_med_n AS (SELECT
COUNT(DISTINCT treat_id) n FROM doc_words_med);
```

Inverteret dokumentfrekvens: Tabellen mono_med_wind

```
CREATE TABLE mono_med_wind AS
SELECT mono_med_tf.treat_id, mono_med_tf.word,
(tf/maxtf)* (LOG(2, n/df)) weight FROM mono_med_tf,
mono_max_med_tf, mono_med_df, mono_n WHERE
mono_med_tf.treat_id=mono_max_med_tf.treat_id AND
mono_med_tf.word=mono_med_df.word;
```

```
CREATE INDEX mono_med_wind_treat_id ON
mono_med_WIND(treat_id);
CREATE INDEX mono_med_wind_word ON
mono_med_WIND(word);
```

Vektormål: Tabellen mono_med_dist

```
CREATE TABLE mono_med_DIST AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance_med(a.treat_id, b.treat_id),0) dist FROM
(SELECT DISTINCT treat_id FROM mono_med_WIND)a,
(SELECT DISTINCT treat_id FROM mono_med_WIND)b
WHERE a.treat_id < b.treat_id;
```

Kun de sidste 10 lægenotater

Klassisk IR undersøgelse med normalisering ved max frekvens og idf-vægtning

Ordtabel: Tabellen doc_words_med10

```
CREATE TABLE doc_words_med10 AS SELECT DISTINCT
t.treat_id, word, t.noteid, fieldid, pos FROM
(SELECT DISTINCT treat_id, noteid,RANK() OVER(
PARTITION BY treat_id ORDER BY noteid DESC) rn FROM
(SELECT DISTINCT treat_id, noteid FROM doc_words WHERE
treat_id IN (SELECT treat_id FROM mono_ud mu UNION
SELECT treat_id FROM mono_skift ms) AND noteid IN
(SELECT DISTINCT noteid FROM noteheader nh WHERE
nh.noteheader like '%æge%') ORDER BY treat_id,
noteid DESC) ORDER BY rn) t, doc_words dw WHERE rn
IN (1,2,3,4,5,6,7,8,9,10) AND t.noteid=dw.noteid;
```

```
CREATE INDEX doc_words_med10_treat_id ON
doc_words_med10(treat_id);
CREATE INDEX doc_words_med10_word ON
doc_words_med10(word);
```

Termfrekvens: Tabellen mono_med10_tf

```
CREATE TABLE mono_med10_tf AS
(SELECT treat_id, word, COUNT(*) tf FROM
doc_words_med10
GROUP BY treat_id, word);
```

```
CREATE INDEX mono_med10_tf_treat_id ON
mono_med10_tf(treat_id);
CREATE INDEX mono_med10_tf_word ON
mono_med10_tf(word);
```

Termnormalisering: Tabellen mono_med10_max_tf

```
CREATE TABLE mono_med10_max_tf AS
(SELECT treat_id, MAX(tf) maxtf FROM
(SELECT * FROM mono_med10_tf)
GROUP BY treat_id);
```

```
CREATE INDEX mono_med10_max_tf_treat_id
ON mono_med10_max_tf(treat_id);
```

Dokumentfrekvens: Tabellen mono_med10_df

```
CREATE TABLE mono_med10_df AS
(SELECT word, COUNT(*) df FROM
(SELECT DISTINCT treat_id, word FROM doc_words_med10)
GROUP BY word);
```

```
CREATE INDEX mono_med10_df_word ON mono_med10_df(word);
```

Elementoptælling: Tabellen mono_med10_n

```
CREATE TABLE mono_med10_n AS
(SELECT COUNT(DISTINCT treat_id) n
FROM doc_words_med10);
```

Inverteret dokumentfrekvens: Tabellen mono_med10_wind

```
CREATE TABLE mono_med10_WIND AS
SELECT mono_med10_tf.treat_id, mono_med10_tf.word,
(tf/maxtf)*(LOG(2, n/df)) weight FROM mono_med10_tf,
mono_med10_max_tf, mono_med10_df, mono_med10_n
WHERE mono_med10_tf.treat_id=mono_med10_max_tf.treat_id
AND mono_med10_tf.word=mono_med10_df.word;
```

```
CREATE INDEX mono_med10_wind_treat_id ON
mono_med10_WIND(treat_id);
CREATE INDEX mono_med10_wind_word ON
mono_med10_WIND(word);
```

Vektormål: Tabellen mono_med10_dist

```
CREATE TABLE mono_med10_DIST AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance_med10(a.treat_id, b.treat_id),0) dist
FROM
(SELECT DISTINCT treat_id FROM mono_med10_WIND)a,
(SELECT DISTINCT treat_id FROM mono_med10_WIND)b
WHERE a.treat_id < b.treat_id;
```

Afstandsberegning: Funktionen distance_med10()

```
create or replace FUNCTION
Distance_med10(treat_id1 INTEGER, treat_id2 INTEGER)
RETURN NUMBER IS

result NUMBER := 0;

BEGIN

WITH ip AS
(SELECT SUM(weight) weight FROM
(SELECT word, SUM(weight) weight FROM
  (SELECT a.word, a.weight* b.weight weight FROM
   (SELECT word, weight FROM mono_med10_WIND WHERE
   treat_id = treat_id1) a,
   (SELECT word, weight FROM mono_med10_WIND WHERE
   treat_id = treat_id2) b
  WHERE a.word=b.word)
GROUP BY word)),
LENGTH AS
(SELECT SQRT(a.len * b.len) len FROM
(SELECT SUM(POWER(weight,2)) len FROM mono_med10_WIND
WHERE treat_id = treat_id1) a,
(SELECT SUM(POWER(weight,2)) len FROM mono_med10_WIND
WHERE treat_id = treat_id2) b)
SELECT weight/len INTO result FROM ip, LENGTH;


RETURN result;
```

Kun de sidste 10% af lægenotaterne

Klassisk IR undersøgelse med normalisering ved max frekvens og idf-vægtning

Ordtabel: Tabellen doc_words_med10perc

```
CREATE TABLE doc_words_med10perc AS SELECT DISTINCT
t.treat_id, word, t.noteid, fieldid, pos
FROM (SELECT DISTINCT treat_id, noteid, RANK() OVER(
PARTITION BY treat_id ORDER BY noteid DESC) rn FROM
(SELECT DISTINCT treat_id, noteid FROM doc_words WHERE
treat_id IN (SELECT treat_id FROM mono_ud mu UNION
SELECT treat_id FROM mono_skift ms) AND noteid IN
(SELECT DISTINCT noteid FROM noteheader nh WHERE
nh.noteheader like '%æge%') ORDER BY treat_id,
noteid DESC) ORDER BY rn ) t, ( SELECT treat_id,
COUNT(DISTINCT noteid) n FROM doc_words GROUP BY
treat_id) tn, doc_words dw WHERE rn <=ROUND(0.1*n)
AND t.noteid=dw.noteid AND t.treat_id = tn.treat_id;
```

```
CREATE INDEX doc_words_med10perc_treat_id ON
doc_words_med10perc(treat_id);
CREATE INDEX doc_words_med10perc_word ON
doc_words_med10perc(word);
```

Termfrekvens: Tabellen mono_med10perc_tf

```
CREATE TABLE mono_med10perc_tf AS
(SELECT treat_id, word, COUNT(*) tf FROM
doc_words_med10perc
GROUP BY treat_id, word);
```

```
CREATE INDEX mono_med10perc_tf_treat_id ON
mono_med10perc_tf(treat_id);
CREATE INDEX mono_med10perc_tf_word ON
mono_med10perc_tf(word);
```

Termnormalisering: Tabellen mono_med10perc_max_tf

```
CREATE TABLE mono_med10perc_max_tf AS
(SELECT treat_id, MAX(tf) maxtf FROM
(SELECT * FROM mono_med10perc_tf)
GROUP BY treat_id);
```

```
CREATE INDEX mono_med10pc_max_tf_treat_id
ON mono_med10perc_max_tf(treat_id);
```

Dokumentfrekvens: Tabellen mono_med10perc_df

```
CREATE TABLE mono_med10perc_df AS
(SELECT word, COUNT(*) df FROM
(SELECT DISTINCT treat_id, word FROM
doc_words_med10perc)
GROUP BY word);
```

```
CREATE INDEX mono_med10perc_df_word
ON mono_med10perc_df(word);
```

Elementoptælling: Tabellen mono_med10perc_n

```
CREATE TABLE mono_med10perc_n AS (SELECT
COUNT(DISTINCT treat_id) n FROM doc_words_med10perc);
```

Inverteret dokumentfrekvens: Tabellen mono_med10perc_wind

```
CREATE TABLE mono_med10perc_WIND AS
SELECT mono_med10perc_tf.treat_id,
mono_med10perc_tf.word, (tf/maxtf)*(LOG(2, n/df))
weight FROM mono_med10perc_tf, mono_med10perc_max_tf,
mono_med10perc_df, mono_med10perc_n WHERE
mono_med10perc_tf.treat_id=
mono_med10perc_max_tf.treat_id AND
mono_med10perc_tf.word=mono_med10perc_df.word;
```

```
CREATE INDEX mono_med10perc_wind_treat_id ON
mono_med10perc_WIND(treat_id);
CREATE INDEX mono_med10perc_wind_word ON
mono_med10perc_WIND(word);
```

Vektormål: Tabellen mono_med10perc_dist

```
CREATE TABLE mono_med10perc_DIST AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance_med10perc(a.treat_id, b.treat_id),0) dist
FROM
(SELECT DISTINCT treat_id FROM mono_med10perc_WIND)a,
(SELECT DISTINCT treat_id FROM mono_med10perc_WIND)b
WHERE a.treat_id < b.treat_id;
```

Afstandsberegning: Funktionen distance_med10perc()

```
create or replace FUNCTION
Distance_med10perc(treat_id1 INTEGER,
treat_id2 INTEGER) RETURN NUMBER IS result
NUMBER := 0;

BEGIN

WITH ip AS
(SELECT SUM(weight) weight FROM
(SELECT word, SUM(weight) weight FROM
  (SELECT a.word, a.weight* b.weight weight FROM
   (SELECT word, weight FROM mono_med10perc_WIND WHERE
   treat_id = treat_id1) a,
   (SELECT word, weight FROM mono_med10perc_WIND WHERE
   treat_id = treat_id2) b
  WHERE a.word=b.word)
GROUP BY word)),
LENGTH AS
(SELECT SQRT(a.len * b.len) len FROM
(SELECT SUM(POWER(weight,2)) len FROM
mono_med10perc_WIND WHERE treat_id = treat_id1) a,
(SELECT SUM(POWER(weight,2)) len FROM
mono_med10perc_WIND WHERE treat_id = treat_id2) b)
SELECT weight/len INTO result FROM ip, LENGTH;


RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

B.2.2 Lemmatisering

Det fulde dokumentsæt

Ordtabel: Tabellen doc_lemma_words

```
CREATE TABLE doc_lemma_words AS
SELECT distinct dr.treat_id,
l.lemma word, aw.noteid, aw.fieldid, aw.pos FROM
drugtreatmentsets ds, drugtregimen dr, all_wdocs aw,
lemma l, notes no, noteheader nh, prescription pr,
drugsprescribed dp WHERE to_date(nh.notetime,
'YYYY-MM-DD') >= dr.begin AND to_date(nh.notetime,
'YYYY-MM-DD') <= dr.end AND no.noteid=nh.noteid AND
no.pid = pr.pid AND dr.treat_id=ds.treat_id AND
ds.treat_id IN (select treat_id from mono_ud mu UNION
SELECT treat_id from mono_skift ms) AND ds.DT_ID=
dp.DT_ID AND dp.PRESCRIPTIONID=pr.PRESCRIPTIONID AND
no.noteid=aw.noteid AND aw.word = l.word;
```

```
DELETE FROM doc_lemma_words WHERE word='<STOPWORD>'
OR word IN (SELECT lemma word from lemma WHERE word
IN(SELECT word from stop_words));
CREATE INDEX doc_lemma_words_treat_id
ON doc_lemma_words(treat_id);
CREATE INDEX doc_lemma_words_word
ON doc_lemma_words(word);
```

Termfrekvens: Tabellen mono_lemma_tf

```
CREATE TABLE mono_lemma_tf AS
(SELECT treat_id, word, COUNT(*) tf
FROM doc_lemma_words GROUP BY treat_id, word);
```

```
CREATE INDEX mono_lemma_tf_treat_id
ON mono_lemma_tf(treat_id);
CREATE INDEX mono_lemma_tf_word
ON mono_lemma_tf(word);
```

Termnormalisering: Tabellen mono_lemma_max_tf

```
CREATE TABLE mono_lemma_max_tf AS (SELECT treat_id,
MAX(tf) maxtf FROM (SELECT * FROM mono_lemma_tf)
GROUP BY treat_id);
```

```
CREATE INDEX mono_lemma_max_tf_treat_id ON
mono_lemma_max_tf(treat_id);
```

Dokumentfrekvens: Tabellen mono_lemma_df

```
CREATE TABLE mono_lemma_df AS (SELECT word,
COUNT(*) df FROM (SELECT DISTINCT treat_id, word
FROM doc_lemma_words) GROUP BY word);
```

```
CREATE INDEX mono_lemma_df_word ON mono_lemma_df(word);
```

Elementoptælling: Tabellen mono_lemma_n

```
CREATE TABLE mono_lemma_n AS (SELECT
COUNT(DISTINCT treat_id) n FROM doc_lemma_words);
```

Normaliseret dokumentfrekvens: Tabellen mono_idflemma_wind

```
CREATE TABLE mono_idflemma_wind AS
SELECT mono_lemma_tf.treat_id, mono_lemma_tf.word,
(tf/maxtf)* (LOG(2, n/df)) weight FROM mono_lemma_tf,
mono_lemma_max_tf, mono_lemma_df, mono_lemma_n
WHERE mono_lemma_tf.treat_id=mono_lemma_max_tf.treat_id
AND mono_lemma_tf.word=mono_lemma_df.word;
```

```
CREATE INDEX mono_idflemma_wind_treat_id
ON mono_idflemma_WIND(treat_id);
CREATE INDEX mono_idflemma_wind_word
ON mono_idflemma_WIND(word);
```

Vektormål: Tabellen mono_idflemma_dist

```
CREATE TABLE mono_idflemma_DIST AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance_idflemma_mono(a.treat_id, b.treat_id),0)
dist FROM
(SELECT DISTINCT treat_id FROM mono_idflemma_WIND)a,
(SELECT DISTINCT treat_id FROM mono_idflemma_WIND)b
WHERE a.treat_id < b.treat_id;
```

Afstandsberegning: Funktionen distance_idflemma_mono()

```
create or replace FUNCTION Distance_lemma_mono(
treat_id1 INTEGER, treat_id2 INTEGER) RETURN NUMBER IS

result NUMBER := 0;

BEGIN

WITH ip AS
(SELECT SUM(weight) weight FROM
(SELECT word, SUM(weight) weight FROM
  (SELECT a.word, a.weight* b.weight weight FROM
   (SELECT word, weight FROM mono_idflemma_WIND
   WHERE treat_id = treat_id1) a,
   (SELECT word, weight FROM mono_idflemma_WIND
   WHERE treat_id = treat_id2) b
  WHERE a.word=b.word)
GROUP BY word)),
LENGTH AS
(SELECT SQRT(a.len * b.len) len FROM
(SELECT SUM(POWER(weight,2)) len
FROM mono_idflemma_WIND
WHERE treat_id = treat_id1) a,
(SELECT SUM(POWER(weight,2)) len
FROM mono_idflemma_WIND
WHERE treat_id = treat_id2) b)
SELECT weight/len INTO result FROM ip, LENGTH;


RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

B.2.3 Uddragning af domænespecifikke termer

Det fulde dokumentsæt

Ordtabel: Tabellen doc_lemmaflip_words

```
CREATE TABLE doc_lemmaflip_words AS
SELECT DISTINCT dr.treat_id, aw.word, aw.noteid,
aw.fieldid, aw.pos FROM drugtreatmentsets ds,
drugtregimen dr, all_wdocs aw, notes no, noteheader nh,
prescription pr, drugsprescribed dp WHERE
TO_DATE(nh.notetime, 'YYYY-MM-DD') >= dr.begin
AND to_date(nh.notetime, 'YYYY-MM-DD') <= dr.end AND
no.noteid=nh.noteid AND no.pid = pr.pid AND
dr.treat_id=ds.treat_id AND ds.treat_id IN (
SELECT treat_id from mono_ud mu UNION
SELECT treat_id from mono_skift ms) AND ds.DT_ID=
dp.DT_ID AND dp.PRESCRIPTIONID=pr.PRESCRIPTIONID AND
no.noteid=aw.noteid;
```

```
DELETE FROM doc_lemmaflip_words WHERE word IN (
SELECT word from lemma_org union select name word
FROM personnames);
CREATE INDEX doc_lemmaflip_words_treat_id
ON doc_lemmaflip_words(treat_id);
CREATE INDEX doc_lemmaflip_words_word
ON doc_lemmaflip_words(word);
```

Termfrekvens: Tabellen mono_lemmaflip_tf

```
CREATE TABLE mono_lemmaflip_tf AS (
SELECT treat_id, word, COUNT(*) tf
FROM doc_lemmaflip_words GROUP BY treat_id, word);
```

```
CREATE INDEX mono_lemmaflip_tf_treat_id
ON mono_lemmaflip_tf(treat_id);
CREATE INDEX mono_lemmaflip_tf_word
ON mono_lemmaflip_tf(word);
```

Termnormalisering: Tabellen mono_lemmaflip_max_tf

```
CREATE TABLE mono_lemmaflip_max_tf AS (
SELECT treat_id, MAX(tf) maxtf FROM (SELECT *
FROM mono_lemmaflip_tf) GROUP BY treat_id);
```

```
CREATE INDEX mono_lemmaflip_mxtf_treat_id ON
mono_lemmaflip_max_tf(treat_id);
```

Dokumentfrekvens: Tabellen mono_lemmaflip_df

```
CREATE TABLE mono_lemmaflip_df AS (
SELECT word, COUNT(*) df FROM (
SELECT DISTINCT treat_id, word
FROM doc_lemmaflip_words) GROUP BY word);
```

```
CREATE INDEX mono_lemmaflip_df_word
ON mono_lemmaflip_df(word);
```

Elementoptælling: Tabellen mono_lemmaflip_n

```
CREATE TABLE mono_lemmaflip_n AS (
SELECT COUNT(DISTINCT treat_id) n
FROM doc_lemmaflip_words);
```

Normaliseret dokumentfrekvens: Tabellen mono_idflemmaflip_wind

```
CREATE TABLE mono_idflemmaflip_wind AS
SELECT mono_lemmaflip_tf.treat_id,
mono_lemmaflip_tf.word, (tf/maxtf)*
(LOG(2, n/df)) weight FROM mono_lemmaflip_tf,
mono_lemmaflip_max_tf, mono_lemmaflip_df,
mono_lemmaflip_n WHERE mono_lemmaflip_tf.treat_id=
mono_lemmaflip_max_tf.treat_id AND
mono_lemmaflip_tf.word=mono_lemmaflip_df.word;
```

```
CREATE INDEX mono_idflemmaflip_wind_treat_id
ON mono_idflemmaflip_WIND(treat_id);
CREATE INDEX mono_idflemmaflip_wind_word
ON mono_idflemmaflip_WIND(word);
```

Vektormål: Tabellen mono_idflemmaflip_dist

```
CREATE TABLE mono_idflemmaflip_DIST AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance_idflemmaflip_mono(
a.treat_id, b.treat_id),0) dist
FROM (SELECT DISTINCT treat_id
FROM mono_idflemmaflip_wind)a,
(SELECT DISTINCT treat_id
```

```
FROM mono_idflemmaflip_wind)b
WHERE a.treat_id < b.treat_id;
```

Afstandsberegning: Funktionen distance_idflemmaflip_mono()

```
CREATE OR REPLACE FUNCTION Distance_idflemmaflip_mono(
treat_id1 INTEGER, treat_id2 INTEGER) RETURN NUMBER IS

result NUMBER := 0;

BEGIN

WITH ip AS
(SELECT SUM(weight) weight FROM
(SELECT word, SUM(weight) weight FROM
  (SELECT a.word, a.weight* b.weight weight FROM
   (SELECT word, weight FROM mono_idflemmaflip_wind
   WHERE treat_id = treat_id1) a,
   (SELECT word, weight FROM mono_idflemmaflip_wind
   WHERE treat_id = treat_id2) b
  WHERE a.word=b.word)
GROUP BY word)),
LENGTH AS
(SELECT SQRT(a.len * b.len) len FROM
(SELECT SUM(POWER(weight,2)) len
FROM mono_idflemmaflip_wind WHERE treat_id =
treat_id1) a,
(SELECT SUM(POWER(weight,2)) len
FROM mono_idflemmaflip_wind WHERE treat_id =
treat_id2) b)
SELECT weight/len INTO result FROM ip, LENGTH;


RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

B.3 Maxtf-vægtning af journalnotater

B.3.1 Uden lemmatisering

Det fulde dokumentsæt

Ved termnormalisering anvendes samme tabelsæt som ved idf-vægtning. Tabellen **mono_wind** der indeholder den vægtede afstand erstattes dog af vægttabellen **mono_wtfmax**.

Vægttabel: Tabellen mono_wtfmax

```
CREATE TABLE mono_tf_wtfmax AS
SELECT mono_tf.treat_id, mono_tf.word, tf/maxtf weight
FROM mono_tf, mono_max_tf WHERE mono_tf.treat_id=
mono_max_tf.treat_id;
```

```
CREATE INDEX mono_wtfmax_treat_id ON
mono_wtfmax(treat_id);
CREATE INDEX mono_wtfmax_word ON
mono_wtfmax(word);
```

Vektormål: Tabellen mono_tf_dist

```
CREATE TABLE mono_tf_dist AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance_wtfmax(a.treat_id, b.treat_id),0) dist
FROM
(SELECT DISTINCT treat_id FROM mono_wtfmax)a,
(SELECT DISTINCT treat_id FROM mono_wtfmax)b
WHERE a.treat_id < b.treat_id;
```

Afstandsberegning: Funktionen distance$_{w}tfmax()$

```
create or replace FUNCTION Distance_wtfmax(treat_id1
INTEGER, treat_id2 INTEGER) RETURN NUMBER IS

result NUMBER := 0;

BEGIN
WITH ip AS
(SELECT SUM(weight) weight FROM
(SELECT word, SUM(weight) weight FROM
  (SELECT a.word, a.weight* b.weight weight FROM
   (SELECT word, weight FROM mono_tf_wtfmax WHERE
```

```
    treat_id = treat_id1) a,
    (SELECT word, weight FROM mono_tf_wtfmax WHERE
    treat_id = treat_id2) b
  WHERE a.word=b.word)
GROUP BY word)),
LENGTH AS
(SELECT SQRT(a.len * b.len) len FROM
(SELECT SUM(POWER(weight,2)) len FROM mono_tf_wtfmax
WHERE
treat_id = treat_id1) a,
(SELECT SUM(POWER(weight,2)) len FROM mono_tf_wtfmax
WHERE
treat_id = treat_id2) b)
SELECT weight/len INTO result FROM ip, LENGTH;

RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

Kun de sidste 10 notater

Vægttabel: Tabellen mono_last10_wtfmax

```
CREATE TABLE mono_last10_wtfmax AS
SELECT mono_last10_tf.treat_id, mono_last10_tf.word,
tf/maxtf weight
FROM mono_last10_tf, mono_last10_max_tf
WHERE mono_last10_tf.treat_id=
mono_last10_max_tf.treat_id;
```

```
CREATE INDEX mono_last10_wtfmax_treat_id ON
mono_last10_wtfmax(treat_id);
CREATE INDEX mono_last10_wtfmax_word ON
mono_last10_wtfmax(word);
```

Vektormål: Tabellen mono_tf_last10_dist

```
CREATE TABLE mono_tf_last10_dist AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance_last10_wtfmax(a.treat_id, b.treat_id),0)
dist FROM
(SELECT DISTINCT treat_id FROM mono_last10_wtfmax)a,
```

```
(SELECT DISTINCT treat_id FROM mono_last10_wtfmax)b
WHERE a.treat_id < b.treat_id;
```

Afstandsberegning: Funktionen distance$_{l}ast10_{w}tfmax()$

```
create or replace FUNCTION
Distance_last10_wtfmax(treat_id1 INTEGER,
treat_id2 INTEGER) RETURN NUMBER IS

result NUMBER := 0;

BEGIN
WITH ip AS
(SELECT SUM(weight) weight FROM
(SELECT word, SUM(weight) weight FROM
  (SELECT a.word, a.weight* b.weight weight FROM
   (SELECT word, weight FROM mono_last10_wtfmax WHERE
   treat_id = treat_id1) a,
   (SELECT word, weight FROM mono_last10_wtfmax WHERE
   treat_id = treat_id2) b
  WHERE a.word=b.word)
GROUP BY word)),
LENGTH AS
(SELECT SQRT(a.len * b.len) len FROM
(SELECT SUM(POWER(weight,2)) len FROM
mono_last10_wtfmax WHERE treat_id = treat_id1) a,
(SELECT SUM(POWER(weight,2)) len FROM
mono_last10_wtfmax WHERE treat_id = treat_id2) b)
SELECT weight/len INTO result FROM ip, LENGTH;

RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

B.3.2 Lemmatisering

Det fulde dokumentsæt

Anvender samme tabelsæt som ved idf-vægtning bortset fra normaliserings og vektormåltabellen.

Normaliseret dokumentfrekvens: Tabellen mono_lemma_wind

```
CREATE TABLE mono_lemma_wind AS
SELECT mono_lemma_tf.treat_id, mono_lemma_tf.word,
tf/maxtf weight FROM mono_lemma_tf, mono_lemma_max_tf
WHERE mono_lemma_tf.treat_id=
mono_lemma_max_tf.treat_id;
```

```
CREATE INDEX mono_lemma_wind_treat_id ON
mono_lemma_WIND(treat_id);
CREATE INDEX mono_lemma_wind_word ON
mono_lemma_WIND(word);
```

Vektormål: Tabellen mono_lemma_dist

```
CREATE TABLE mono_lemma_DIST AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance_lemma_mono(a.treat_id, b.treat_id),0)
dist FROM
(SELECT DISTINCT treat_id FROM mono_lemma_WIND)a,
(SELECT DISTINCT treat_id FROM mono_lemma_WIND)b
WHERE a.treat_id < b.treat_id;
```

Afstandsberegning: Funktionen distance_lemma_mono()

```
CREATE OR REPLACE FUNCTION
Distance_lemma_mono(treat_id1 INTEGER,
treat_id2 INTEGER) RETURN NUMBER IS

result NUMBER := 0;

BEGIN

WITH ip AS
(SELECT SUM(weight) weight FROM
(SELECT word, SUM(weight) weight FROM
  (SELECT a.word, a.weight* b.weight weight FROM
   (SELECT word, weight FROM mono_lemma_WIND WHERE
```

```
    treat_id = treat_id1) a,
    (SELECT word, weight FROM mono_lemma_WIND WHERE
    treat_id = treat_id2) b
  WHERE a.word=b.word)
GROUP BY word)),
LENGTH AS
(SELECT SQRT(a.len * b.len) len FROM
(SELECT SUM(POWER(weight,2)) len FROM mono_lemma_WIND
WHERE treat_id = treat_id1) a,
(SELECT SUM(POWER(weight,2)) len FROM mono_lemma_WIND
WHERE treat_id = treat_id2) b)
SELECT weight/len INTO result FROM ip, LENGTH;


RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

De sidste 10 notater

Ordtabel: Tabellen doc_lemma_words_last10

```
CREATE TABLE doc_lemma_words_last10 AS SELECT DISTINCT
t.treat_id, word, t.noteid, fieldid, pos FROM (SELECT
distinct treat_id, noteid, RANK() OVER(partition by
treat_id ORDER BY noteid desc) rn
FROM (SELECT DISTINCT treat_id, noteid
FROM doc_lemma_words WHERE treat_id IN (
SELECT treat_id FROM mono_ud mu UNION SELECT treat_id
FROM mono_skift ms) ORDER BY treat_id, noteid desc)
ORDER BY rn) t, doc_lemma_words dw WHERE rn
in (1,2,3,4,5,6,7,8,9,10) and t.noteid=dw.noteid;
```

Termfrekvens: Tabellen mono_lemma_last10_tf

```
CREATE TABLE mono_lemma_last10_tf AS
(SELECT treat_id, word, COUNT(*) tf FROM
doc_lemma_words_last10 GROUP BY treat_id, word);
```

```
CREATE INDEX mono_lemlast10_tf_treat_id ON
mono_lemma_last10_tf(treat_id);
CREATE INDEX mono_lemlast10_tf_word ON
mono_lemma_last10_tf(word);
```

Termnormalisering: Tabellen mono_lemma_last10_max_tf

```
CREATE TABLE mono_lemma_last10_max_tf AS
(SELECT treat_id, MAX(tf) maxtf FROM
(SELECT * FROM mono_lemma_last10_tf)
GROUP BY treat_id);
```

```
CREATE INDEX mono_lemlast10_max_tf_treat_id ON
mono_lemma_last10_max_tf(treat_id);
```

Dokumentfrekvens: Tabellen mono_lemma_last10_df

```
CREATE TABLE mono_lemma_last10_df AS (SELECT word,
COUNT(*) df FROM (SELECT DISTINCT treat_id, word FROM
doc_lemma_words_last10) GROUP BY word);
```

```
CREATE INDEX mono_lemlast10_df_word
ON mono_lemma_last10_df(word);
```

Elementoptælling: Tabellen mono_lemma_last10_n

```
CREATE TABLE mono_lemma_last10_n AS (SELECT
COUNT(DISTINCT treat_id) n FROM
doc_lemma_words_last10);
```

Dokumentfrekvens: Tabellen mono_lemma_last10_wind

```
CREATE TABLE mono_lemma_last10_WIND AS SELECT
mono_lemma_last10_tf.treat_id,
mono_lemma_last10_tf.word, tf/maxtf weight FROM
mono_lemma_last10_tf, mono_lemma_last10_max_tf
WHERE mono_lemma_last10_tf.treat_id=
mono_lemma_last10_max_tf.treat_id;
```

```
CREATE INDEX mono_lemlast10_wind_treat_id ON
mono_lemma_last10_WIND(treat_id);
CREATE INDEX mono_lemlast10_wind_word ON
mono_lemma_last10_WIND(word);
```

Vektormål: Tabellen mono_lemma_last10_dist

```
CREATE TABLE mono_lemma_last10_DIST AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance_lemma_last10(a.treat_id, b.treat_id),0)
dist FROM
```

```
(SELECT DISTINCT treat_id
FROM mono_lemma_last10_WIND)a,
(SELECT DISTINCT treat_id
FROM mono_lemma_last10_WIND)b
WHERE a.treat_id < b.treat_id;
```

Afstandsberegning: Funktionen distance_lemma_last10()

```
CREATE OR REPLACE FUNCTION
Distance_lemma_last10(treat_id1 INTEGER,
treat_id2 INTEGER) RETURN NUMBER IS

result NUMBER := 0;

BEGIN

WITH ip AS
(SELECT SUM(weight) weight FROM
(SELECT word, SUM(weight) weight FROM
  (SELECT a.word, a.weight* b.weight weight FROM
   (SELECT word, weight FROM mono_lemma_last10_WIND
   WHERE treat_id = treat_id1) a,
   (SELECT word, weight FROM mono_lemma_last10_WIND
   WHERE treat_id = treat_id2) b
  WHERE a.word=b.word)
GROUP BY word)),
LENGTH AS
(SELECT SQRT(a.len * b.len) len FROM
(SELECT SUM(POWER(weight,2)) len FROM
mono_lemma_last10_WIND WHERE treat_id = treat_id1) a,
(SELECT SUM(POWER(weight,2)) len FROM
mono_lemma_last10_WIND WHERE treat_id = treat_id2) b)
SELECT weight/len INTO result FROM ip, LENGTH;

RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

B.3.3 Lemmatisering med øget vægt på vigtige kliniske termer

De sidste 10 notater

Vægtjustering af dokumentfrevensen: Tabellen mono_lemma_last10boost_wind

```
CREATE TABLE mono_lemma_last10boost_wind AS
SELECT * FROM mono_lemma_last10_wind;
UPDATE mono_lemma_last10boost_wind SET weight=20000
WHERE word IN (SELECT lemma word FROM lemma
WHERE word IN (SELECT word FROM w_kdj));
```

```
CREATE INDEX mono_lemlast10bo_wind_treat_id
ON mono_lemma_last10boost_WIND(treat_id);
CREATE INDEX mono_lemlast10bo_wind_word
ON mono_lemma_last10boost_WIND(word);
```

Vektormål: Tabellen mono_lemma_last10boost_dist

```
CREATE TABLE mono_lemma_last10boost_DIST AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2, NVL(
Distance_lemma_last10boost(a.treat_id, b.treat_id),0)
dist FROM (SELECT DISTINCT treat_id
FROM mono_lemma_last10boost_WIND)a,
(SELECT DISTINCT treat_id
FROM mono_lemma_last10boost_WIND)b
WHERE a.treat_id < b.treat_id;
```

B.3.4 Uddragning af domænespecifikke termer

Det fulde dokumentsæt

Anvender samme tabelsæt som ved idf-vægtning bortset fra normaliserings og vektormåltabellen.

Normaliseret dokumentfrekvens: Tabellen mono_lemmaflip_wind

```
CREATE TABLE mono_lemmaflip_wind AS
SELECT mono_lemmaflip_tf.treat_id,
mono_lemmaflip_tf.word, tf/maxtf weight
FROM mono_lemmaflip_tf, mono_lemmaflip_max_tf
WHERE mono_lemmaflip_tf.treat_id=
mono_lemmaflip_max_tf.treat_id;
```

```
CREATE INDEX mono_lemmaflip_wind_treat_id
ON mono_lemmaflip_WIND(treat_id);
CREATE INDEX mono_lemmaflip_wind_word
ON mono_lemmaflip_WIND(word);
```

Vektormål: Tabellen mono_lemmaflip_dist

```
CREATE TABLE mono_lemmaflip_DIST AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance_lemmaflip_mono(a.treat_id, b.treat_id),0)
dist FROM (SELECT DISTINCT treat_id
FROM mono_lemmaflip_wind)a, (SELECT DISTINCT treat_id
FROM mono_lemmaflip_wind)b
WHERE a.treat_id < b.treat_id;
```

Afstandsberegning: Funktionen distance_lemmaflip_mono()

```
CREATE OR REPLACE FUNCTION Distance_lemmaflip_mono(
treat_id1 INTEGER, treat_id2 INTEGER) RETURN NUMBER IS

result NUMBER := 0;

BEGIN

WITH ip AS
(SELECT SUM(weight) weight FROM
(SELECT word, SUM(weight) weight FROM
  (SELECT a.word, a.weight* b.weight weight FROM
   (SELECT word, weight FROM mono_lemmaflip_wind
```

```
    WHERE treat_id = treat_id1) a,
    (SELECT word, weight FROM mono_lemmaflip_wind
    WHERE treat_id = treat_id2) b
  WHERE a.word=b.word)
GROUP BY word)),
LENGTH AS
(SELECT SQRT(a.len * b.len) len FROM
(SELECT SUM(POWER(weight,2)) len
FROM mono_lemmaflip_wind WHERE treat_id = treat_id1) a,
(SELECT SUM(POWER(weight,2)) len
FROM mono_lemmaflip_wind WHERE treat_id = treat_id2) b)
SELECT weight/len INTO result FROM ip, LENGTH;


RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

B.4 Jaccard similaritetsmål for journalnotater

B.4.1 Uden lemmatisering

Det fulde dokumentsæt

Ved termnormalisering anvendes samme tabelsæt som ved idf-vægtning.

Vektormål: Tabellen mono_jaccard_dist

```
CREATE TABLE mono_jaccard_dist AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(jaccard(a.treat_id, b.treat_id),0) dist FROM
(SELECT DISTINCT treat_id FROM mono_tf)a,
(SELECT DISTINCT treat_id FROM mono_tf)b
WHERE a.treat_id < b.treat_id;
```

Afstandsberegning: Funktionen jaccard()

```
create or replace FUNCTION Jaccard(treat_id1 INTEGER,
treat_id2 INTEGER) RETURN NUMBER IS

result NUMBER := 0;

BEGIN

select top/but into result from
(select sum(least(a.tf, b.tf)) top from
    (SELECT word, tf FROM mono_tf WHERE treat_id =
    treat_id1) a,
    (SELECT word, tf FROM mono_tf WHERE treat_id =
    treat_id2) b
where a.word = b.word) aa,
(select sum(tf) but from
(select word, max(tf) tf from
(SELECT word, tf FROM mono_tf WHERE treat_id =
treat_id1
union all
SELECT word, tf FROM mono_tf WHERE treat_id =
treat_id2)
group by word)) bb;

RETURN result;

EXCEPTION WHEN OTHERS THEN
```

```
RETURN 0;
END;
```

B.5 Patientvariable som similaritetsmål

B.5.1 Alder

Similaritetstabel: Tabellen age_dist

```
CREATE TABLE age_dist AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance_age(a.treat_id, b.treat_id),0) dist FROM
(SELECT treat_id FROM mono_ud UNION SELECT treat_id
FROM mono_skift)a,
(SELECT treat_id FROM mono_ud UNION SELECT treat_id
FROM mono_skift)b
WHERE a.treat_id < b.treat_id;
```

Aldersbestemmelse: Funktionen age() Funktionen returnerer patientens alder ved behandlingsperiodens start.

```
CREATE OR REPLACE FUNCTION Age(treat_id1 INTEGER)
RETURN NUMBER IS

result NUMBER := 0;

BEGIN

WITH age_calc AS
(SELECT ROUND((begin-to_date(SUBSTR(
pe.birthdate, 1, 10), 'yyyy-mm-dd'))/365) age
FROM drugtregimen dtr, person pe, ipe, ipes
WHERE dtr.groupid=ipes.groupid
AND ipes.ipeid = ipe.ipeid AND ipe.pid = pe.pid AND
treat_id = treat_id1)
SELECT age INTO result FROM age_calc;

RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

Alderslighed: Funktionen Distance_age() Funktionen returnerer procentdelen den laveste alder udgør af den højeste alder.

```
CREATE OR REPLACE FUNCTION Distance_Age(
treat_id1 INTEGER, treat_id2 INTEGER) RETURN NUMBER IS
```

```
result NUMBER := 0;

BEGIN

WITH min_age AS
(SELECT AGE(treat_id) age FROM drugtregimen
WHERE (treat_id =treat_id1 OR treat_id = treat_id2) AND
AGE(treat_id) IN (SELECT MIN(age) FROM (SELECT
AGE(treat_id) age from drugtregimen
WHERE treat_id =treat_id2 UNION
SELECT AGE(treat_id) age FROM drugtregimen
WHERE treat_id =treat_id1))), max_age AS
(SELECT AGE(treat_id) age FROM drugtregimen
WHERE (treat_id =treat_id1 OR treat_id = treat_id2)
AND AGE(treat_id) IN(SELECT MAX(age) FROM (SELECT
AGE(treat_id) age FROM drugtregimen
WHERE treat_id =treat_id2 UNION SELECT AGE(treat_id) age
FROM drugtregimen where treat_id =treat_id1)))
SELECT min_age.age/max_age.age INTO result
FROM min_age, max_age;


RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

B.5.2 Køn

Similaritetstabel: Tabellen sex_dist

```
CREATE TABLE sex_dist AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance_sex(a.treat_id, b.treat_id),0) dist FROM
(SELECT treat_id FROM mono_ud UNION SELECT treat_id
FROM mono_skift)a,
(SELECT treat_id FROM mono_ud UNION SELECT treat_id
FROM mono_skift)b WHERE a.treat_id < b.treat_id;
```

Kønsbestemmelse: Funktionen sex() Funktionen returnerer patientens køn.

```
CREATE OR REPLACE FUNCTION
```

```
Sex(treat_id1 INTEGER) RETURN NUMBER IS

result NUMBER := 0;

BEGIN

WITH gender AS
(SELECT
  (CASE
    WHEN sex like 'M' THEN 1
    WHEN sex like 'F' THEN 2
    ELSE 0
  END ) gender
FROM person pe, drugtregimen dtr, ipes, ipe
WHERE dtr.groupid = ipes.groupid
AND ipes.ipeid = ipe.ipeid AND ipe.pid = pe.pid
AND treat_id = treat_id1)
SELECT gender INTO result FROM gender;

RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

Kønslighed: Funktionen Distance_sex() Funktionen returnerer tallet 1 ved samme køn ellers 0.

```
CREATE OR REPLACE FUNCTION Distance_sex(
treat_id1 INTEGER, treat_id2 INTEGER) RETURN NUMBER IS

result NUMBER := 0;

BEGIN

WITH sex AS
(SELECT
  (CASE
    WHEN a.gender = b.gender THEN 1
    ELSE 0
  END ) same_sex
FROM
(SELECT sex(treat_id) gender FROM drugtregimen
WHERE treat_id = treat_id1) a,
```

```
(SELECT sex(treat_id) gender FROM drugtregimen
WHERE treat_id = treat_id2) b)
SELECT same_sex INTO result FROM sex;

RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

B.5.3 Afdeling

Similaritetstabel: Tabellen ward_dist

```
CREATE TABLE ward_dist AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance_ward(a.treat_id, b.treat_id),0) dist FROM
(SELECT treat_id FROM mono_ud UNION SELECT treat_id
FROM mono_skift)a,
(SELECT treat_id FROM mono_ud UNION SELECT treat_id
FROM mono_skift)b WHERE a.treat_id < b.treat_id;
```

Afdelingsbestemmelse: Funktionen ward() Funktionen returnerer en identifikator for den tilknyttede afdeling.

```
CREATE OR REPLACE FUNCTION Ward(treat_id1 INTEGER)
RETURN NUMBER IS result NUMBER := 0;

BEGIN

WITH afdeling AS
(SELECT
  (CASE
    WHEN name like 'Afdeling P%' THEN 1
    WHEN name like 'Afdeling M%' THEN 2
    WHEN name like 'Afdeling U%' THEN 3
    WHEN name like 'Afdeling R%' THEN 4
    WHEN name like 'Afdeling L%' THEN 5
    ELSE 0
  END ) ward
FROM clinic cli, ipe, ipes, drugtregimen dtr
WHERE dtr.groupid = ipes.groupid AND ipe.ipeid =
ipes.ipeid AND ipe.clinicid = cli.clinicid
AND BEGIN < to_date(SUBSTR(ipe.dischargedt, 1, 10),
```

```
'yyyy-mm-dd') AND END > to_date(SUBSTR(
ipe.registrationdt, 1, 10), 'yyyy-mm-dd')
AND treat_id = treat_id1) SELECT ward INTO result
FROM afdeling;

RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

Afdelingslighed: Funktionen Distance_ward() Funktionen returnerer tallet 1 ved samme køn ellers 0.

```
CREATE OR REPLACE FUNCTION Distance_ward(
treat_id1 INTEGER, treat_id2 INTEGER) RETURN NUMBER IS

result NUMBER := 0;

BEGIN

WITH afdeling AS
(SELECT
  (CASE
    WHEN a.ward = b.ward THEN 1
    ELSE 0
  END ) same_ward
FROM
(SELECT ward(treat_id) ward FROM drugtregimen
WHERE treat_id = treat_id1) a,
(SELECT ward(treat_id) ward FROM drugtregimen
WHERE treat_id = treat_id2) b)
SELECT same_ward INTO result FROM afdeling;

RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

B.5.4 Diagnose (ukategoriseret hoved-/bidiagnose)

Similaritetstabel: Tabellen diagnosis_ukat_dist

```
CREATE TABLE diagnosis_ukat_dist AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance_diagnosis_ukat(a.treat_id, b.treat_id),0)
dist FROM
(SELECT treat_id FROM mono_ud UNION SELECT treat_id
FROM mono_skift)a,
(SELECT treat_id FROM mono_ud UNION SELECT treat_id
FROM mono_skift)b WHERE a.treat_id < b.treat_id;
```

Diagnoselighed: Funktionen Distance_diagnosis_ukat() Funktionen returnerer tallet mellem 0 og 1 afhængig af lighed i diagnosespektret.

```
CREATE OR REPLACE FUNCTION DISTANCE_DIAGNOSIS_UKAT(
treat_id1 INTEGER, treat_id2 INTEGER) RETURN NUMBER IS

result NUMBER := 0;

BEGIN

with top AS
(select max(antal) antal from (
select antal from (
(SELECT count(F_DIAGNOSE) antal
FROM DIAGNOSE_UKATEGORISERET WHERE treat_id = treat_id1
UNION SELECT count(F_DIAGNOSE) antal
FROM DIAGNOSE_UKATEGORISERET
WHERE treat_id = treat_id2 )))), incommon AS
(select count(F_DIAGNOSE) antal from (
select F_DIAGNOSE FROM DIAGNOSE_UKATEGORISERET
WHERE treat_id = treat_id1 intersect
SELECT F_DIAGNOSE FROM DIAGNOSE_UKATEGORISERET
WHERE treat_id = treat_id2))
SELECT incommon.antal/top.antal INTO result
FROM top, incommon;


RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

B.5.5 Diagnose (kategoriseret)

Similaritetstabel: Tabellen diagnosis_kat_dist

```
CREATE TABLE diagnosis_kat_dist AS
SELECT a.treat_id treat_id1, b.treat_id treat_id2,
NVL(Distance_diagnosis_kat(a.treat_id, b.treat_id),0)
dist FROM (SELECT DISTINCT treat_id
FROM diagnose_kategoriseret)a,
(SELECT DISTINCT treat_id
FROM diagnose_kategoriseret)b
WHERE a.treat_id < b.treat_id;
```

Diagnoselighed: Funktionen Distance_diagnosis_kat() Funktionen returnerer tallet mellem 0 og 1 afhængig af lighed i diagnosespektret.

```
CREATE OR REPLACE FUNCTION DISTANCE_DIAGNOSIS_KAT(
treat_id1 INTEGER, treat_id2 INTEGER) RETURN NUMBER IS

result NUMBER := 0;

BEGIN

with hoved AS (select max(antal) antal from (
select antal from ((SELECT count(F_DIAGNOSE) antal
FROM DIAGNOSE_KATEGORISERET WHERE treat_id = treat_id1
AND hoved_bi = 1 union SELECT count(F_DIAGNOSE) antal
FROM DIAGNOSE_KATEGORISERET WHERE treat_id = treat_id2
AND hoved_bi = 1)))), h_incommon AS (
SELECT COUNT(F_DIAGNOSE) antal FROM (
SELECT F_DIAGNOSE FROM DIAGNOSE_KATEGORISERET
WHERE treat_id = treat_id1 AND hoved_bi = 1 intersect
SELECT F_DIAGNOSE FROM DIAGNOSE_KATEGORISERET
WHERE treat_id = treat_id2 AND hoved_bi = 1
)),
bi AS
(SELECT MAX(antal) antal FROM (
SELECT antal FROM (
(SELECT COUNT(F_DIAGNOSE) antal
FROM DIAGNOSE_KATEGORISERET WHERE treat_id = treat_id1
AND hoved_bi = 2 UNION
SELECT COUNT(F_DIAGNOSE) antal
FROM DIAGNOSE_KATEGORISERET WHERE treat_id = treat_id2
AND hoved_bi = 2)))),b_incommon AS
(SELECT COUNT(F_DIAGNOSE) antal FROM (
```

```
select F_DIAGNOSE FROM DIAGNOSE_KATEGORISERET
WHERE treat_id = treat_id1 AND hoved_bi = 2 intersect
SELECT F_DIAGNOSE FROM DIAGNOSE_KATEGORISERET
WHERE treat_id = treat_id2 AND hoved_bi = 2
)), max_antal AS (SELECT MAX(antal) antal FROM (
SELECT antal FROM (SELECT COUNT(*) antal from (
SELECT DISTINCT treat_id, hoved_bi
FROM DIAGNOSE_KATEGORISERET
WHERE treat_id = treat_id1) UNION
SELECT COUNT(*) antal FROM (SELECT DISTINCT treat_id,
hoved_bi FROM DIAGNOSE_KATEGORISERET
WHERE treat_id = treat_id2))))
SELECT ((b_incommon.antal/bi.antal)+
(h_incommon.antal/hoved.antal))/max_antal.antal
INTO result
FROM bi, b_incommon, hoved, h_incommon, max_antal;

RETURN result;

EXCEPTION WHEN OTHERS THEN
RETURN 0;
END;
```

B.6 Koblede patientvariable

B.6.1 Alder og køn

Similaritetstabel: Tabellen kobling_age_sex

```
CREATE TABLE kobling_age_sex AS
SELECT a.treat_id1, a.treat_id2, a.dist * b.dist dist
FROM age_dist a, sex_dist b WHERE a.treat_id1=b.treat_id1
AND a.treat_id2=b.treat_id2;
```

B.6.2 Alder og afdeling

Similaritetstabel: Tabellen kobling_age_ward

```
CREATE TABLE kobling_age_ward AS
SELECT a.treat_id1, a.treat_id2, a.dist * b.dist dist FROM
age_dist a, ward_dist b WHERE a.treat_id1=b.treat_id1 AND
a.treat_id2=b.treat_id2;
```

B.6.3 Køn og afdeling

Similaritetstabel: Tabellen kobling_sex_ward

```
CREATE TABLE kobling_sex_ward AS
SELECT a.treat_id1, a.treat_id2, a.dist * b.dist dist FROM
sex_dist a, ward_dist b WHERE a.treat_id1=b.treat_id1 AND
a.treat_id2=b.treat_id2;
```

B.6.4 Alder, køn og afdeling

Similaritetstabel: Tabellen kobling_age_sex_ward

```
CREATE TABLE kobling_age_sex_ward AS
SELECT a.treat_id1, a.treat_id2, a.dist * b.dist dist FROM
kobling_age_sex a, ward_dist b WHERE a.treat_id1=b.treat_id1
AND a.treat_id2=b.treat_id2;
```

B.6.5 Alder, køn og diagnose (ukategoriseret)

Similaritetstabel: Tabellen kobling_age_sex_diag_u

```
CREATE TABLE kobling_age_sex_diag_u AS
SELECT a.treat_id1, a.treat_id2, a.dist * b.dist dist FROM
kobling_age_sex a, diagnosis_ukat_dist b
WHERE a.treat_id1=b.treat_id1 AND a.treat_id2=b.treat_id2;
```

B.6.6 Alder, køn og diagnose (kategoriseret)

Similaritetstabel: Tabellen kobling_age_sex_diag_k

```
CREATE TABLE kobling_age_sex_diag_k AS
SELECT a.treat_id1, a.treat_id2, a.dist * b.dist dist FROM
kobling_age_sex a, diagnosis_kat_dist b
WHERE a.treat_id1=b.treat_id1 AND a.treat_id2=b.treat_id2;
```

B.6.7 Alder, afdeling og diagnose (ukategoriseret)

Similaritetstabel: Tabellen kobling_age_ward_diag_u

```
CREATE TABLE kobling_age_ward_diag_u AS
SELECT a.treat_id1, a.treat_id2, a.dist * b.dist dist FROM
kobling_age_ward a, diagnosis_ukat_dist b
WHERE a.treat_id1=b.treat_id1 AND a.treat_id2=b.treat_id2;
```

B.6.8 Alder, afdeling og diagnose (kategoriseret)

Similaritetstabel: Tabellen kobling_age_ward_diag_k

```
CREATE TABLE kobling_age_ward_diag_k AS
SELECT a.treat_id1, a.treat_id2, a.dist * b.dist dist FROM
kobling_age_ward a, diagnosis_kat_dist b
WHERE a.treat_id1=b.treat_id1 AND a.treat_id2=b.treat_id2;
```

B.6.9 Køn, afdeling og diagnose (ukategoriseret)

Similaritetstabel: Tabellen kobling_sex_ward_diag_u

```
CREATE TABLE kobling_sex_ward_diag_u AS
SELECT a.treat_id1, a.treat_id2, a.dist * b.dist dist FROM
kobling_sex_ward a, diagnosis_ukat_dist b
WHERE a.treat_id1=b.treat_id1 AND a.treat_id2=b.treat_id2;
```

B.6.10 Køn, afdeling og diagnose (kategoriseret)

Similaritetstabel: Tabellen kobling_sex_ward_diag_k

```
CREATE TABLE kobling_sex_ward_diag_k AS
SELECT a.treat_id1, a.treat_id2, a.dist * b.dist dist FROM
kobling_sex_ward a, diagnosis_kat_dist b
WHERE a.treat_id1=b.treat_id1 AND a.treat_id2=b.treat_id2;
```

B.6.11 Alder, køn, afdeling og diagnose (ukategoriseret)

Similaritetstabel: Tabellen kobling_age_sex_ward_diag_u

```
CREATE TABLE kobling_age_sex_ward_diag_u AS
SELECT a.treat_id1, a.treat_id2, a.dist * b.dist dist FROM
kobling_age_sex_ward a, diagnosis_ukat_dist b
WHERE a.treat_id1=b.treat_id1 AND a.treat_id2=b.treat_id2;
```

B.6.12 Alder, køn, afdeling og diagnose (kategoriseret)

Similaritetstabel: Tabellen kobling_age_sex_ward_diag_k

```
CREATE TABLE kobling_age_sex_ward_diag_k AS
SELECT a.treat_id1, a.treat_id2, a.dist * b.dist dist FROM
kobling_age_sex_ward a, diagnosis_kat_dist b
WHERE a.treat_id1=b.treat_id1 AND a.treat_id2=b.treat_id2;
```

B.7 Kobling krydsede variable

B.7.1 Alder, afdeling og idf-vægtede sidste 10% af samtlige notater uden lemmatisering

Similaritetstabel: Tabellen kobling_age_ward_idf_last10

```
CREATE TABLE kobling_age_ward_idf_last10 AS
SELECT a.treat_id1, a.treat_id2, a.dist * b.dist dist FROM
kobling_age_ward a, mono_last10perc_DIST b
WHERE a.treat_id1=b.treat_id1 AND a.treat_id2=b.treat_id2;
```

Bilag C

Clusterfordelinger

C.1 Idf-vægtning

C.1.1 Uden lemmatisering

Det fulde dokumentsæt

Clusterfordelingen for det fulde dokumentsæt er angivet i tabel C.1.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	56.557377%	43.442623%	122
Cluster 2	66.81035%	33.189655%	232

Tabel C.1: Clusterfordeling ved 2 clusters med idf-vægtning, uden lemmatisering over det fulde dokumentsæt.

Kun de sidste 10 notater

Clusterfordelingen for de sidste 10 notater er angivet i tabel C.2.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	63.636364%	36.363636%	286
Cluster 2	61.764706%	38.235294%	68

Tabel C.2: Clusterfordeling ved 2 clusters med idf-vægtning, uden lemmatisering over de sidste 10 notater.

Kun de sidste 10% af samtlige notater

Clusterfordelingen for de sidste 10% af samtlige notater er angivet i tabel C.3.

Kun lægenotater

Clusterfordelingen for lægenotater alene er angivet i tabel C.4.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	74.3421%	25.657894%	152
Cluster 2	54.950497%	45.049503%	202

Tabel C.3: Clusterfordeling ved 2 clusters med idf-vægtning, uden lemmatisering over de sidste 10% af notaterne.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	50.0%	50.0%	2
Cluster 2	63.352272%	36.647728%	352

Tabel C.4: Clusterfordeling ved 2 clusters med idf-vægtning, uden lemmatisering over lægenotaterne.

Kun de sidste 10 lægenotater

Clusterfordelingen for de sidste 10 lægenotater er angivet i tabel C.5.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	50.0%	50.0%	2
Cluster 2	63.352272%	36.647728%	352

Tabel C.5: Clusterfordeling ved 2 clusters med idf-vægtning, uden lemmatisering over de sidste 10 lægenotater.

Kun de sidste 10% af lægenotaterne

Clusterfordelingen for de sidste 10% af lægenotaterne er angivet i tabel C.6.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	50.0%	50.0%	2
Cluster 2	63.352272%	36.647728%	352

Tabel C.6: Clusterfordeling ved 2 clusters med idf-vægtning, uden lemmatisering over de sidste 10% af lægenotaterne.

C.1.2 Lemmatisering

Det fulde dokumentsæt

Clusterfordelingen for det fulde dokumentsæt er angivet i tabel C.7.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	64.31925%	35.68075%	213
Cluster 2	61.70213%	38.29787%	141

Tabel C.7: Clusterfordeling ved 2 clusters med idf-vægtning og lemmatisering over det fulde dokumentsæt.

C.1.3 Uddragning af domænespecifikke termer

Det fulde dokumentsæt

Clusterfordelingen for det fulde dokumentsæt er angivet i tabel C.8.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	56.756756%	43.243244%	111
Cluster 2	66.25514%	33.744858%	243

Tabel C.8: Clusterfordeling ved 2 clusters med idf-vægtning og uddragning af domænespecifikke termer over det fulde dokumentsæt.

C.2 Maxtf-vægtning

C.2.1 Uden lemmatisering

Det fulde dokumentsæt

Clusterfordelingen for det fulde dokumentsæt er angivet i tabel C.9.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	63.142857%	36.857143%	350
Cluster 2	75.0%	25.0%	4

Tabel C.9: Clusterfordeling ved 2 clusters med maxtf-vægtning, uden lemmatisering over det fulde dokumentsæt.

Kun de sidste 10 notater

Clusterfordelingen for de sidste 10 notater er angivet i tabel C.10.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	58.82353%	41.17647%	255
Cluster 2	74.747475%	25.252525%	99

Tabel C.10: Clusterfordeling ved 2 clusters med maxtf-vægtning, uden lemmatisering over de sidste 10 notater.

C.2.2 Lemmatisering

Det fulde dokumentsæt

Clusterfordelingen for det fulde dokumentsæt er angivet i tabel C.11.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	72.94118%	27.058823%	85
Cluster 2	60.22305%	39.77695%	269

Tabel C.11: Clusterfordeling ved 2 clusters med maxtf-vægtning og lemmatisering over det fulde dokumentsæt.

Kun de sidste 10 notater

Clusterfordelingen for de sidste 10 notater er angivet i tabel C.12.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	64.86487%	35.135136%	259
Cluster 2	58.94737%	41.05263%	95

Tabel C.12: Clusterfordeling ved 2 clusters med maxtf-vægtning og lemmatisering over de sidste 10 notater.

C.2.3 Lemmatisering med øget vægt på vigtige kliniske termer

De sidste 10 notater

Clusterfordelingen for de sidste 10 notater er angivet i tabel C.13.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	63.384617%	36.615383%	325
Cluster 2	62.068966%	37.931034%	29

Tabel C.13: Clusterfordeling ved 2 clusters med maxtf-vægtning og lemmatisering med øget vægt på vigtige kliniske termer over de sidste 10 notater.

C.2.4 Uddragning af domænespecifikke termer

Det fulde dokumentsæt

Clusterfordelingen for det fulde dokumentsæt er angivet i tabel C.14.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	58.86525%	41.13475%	141
Cluster 2	66.19718%	33.80282%	213

Tabel C.14: Clusterfordeling ved 2 clusters med maxtf-vægtning og uddragning af domænespecifikke termer over det fulde dokumentsæt.

C.3 Jaccard similaritetsmål for journalnotater

C.3.1 Uden lemmatisering

Det fulde dokumentsæt

Clusterfordelingen for det fulde dokumentsæt er angivet i tabel C.15.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	66.32124%	33.678757%	193
Cluster 2	59.62733%	40.37267%	161

Tabel C.15: Clusterfordeling ved 2 clusters med Jaccard som similaritetsmål, uden lemmatisering over det fulde dokumentsæt.

C.4 Patientvariable som similaritetsmål

C.4.1 Alder

Clusterfordelingen med alder som similaritetsmål er angivet i tabel C.16.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	50.0%	50.0%	20
Cluster 2	64.07185%	35.928143%	334

Tabel C.16: Clusterfordeling ved 2 clusters med alder som similaritetsmål.

C.4.2 Køn

Clusterfordelingen med køn som similaritetsmål er angivet i tabel C.17.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	57.042255%	42.957745%	142
Cluster 2	67.45283%	32.54717%	212

Tabel C.17: Clusterfordeling ved 2 clusters med køn som similaritetsmål.

C.4.3 Afdeling

Clusterfordelingen med afdelingsnavn som similaritetsmål er angivet i tabel C.18.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	68.776375%	31.223629%	237
Cluster 2	52.136753%	47.863247%	117

Tabel C.18: Clusterfordeling ved 2 clusters med afdeling som similaritetsmål.

C.4.4 Diagnose (ukategoriseret hoved-/bidiagnose)

Clusterfordelingen med diagnosebetegnelse som similaritetsmål er angivet i tabel C.19.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	70.754715%	29.245283%	106
Cluster 2	60.080647%	39.919353%	248

Tabel C.19: Clusterfordeling ved 2 clusters med diagnosebetegnelse som similaritetsmål.

C.4.5 Diagnose (kategoriseret)

Clusterfordelingen med diagnosebetegnelse for hoved- og bidiagnose som similaritetsmål er angivet i tabel C.20.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	62.5%	37.5%	344
Cluster 2	90.0%	10.0%	10

Tabel C.20: Clusterfordeling ved 2 clusters med diagnosebetegnelse for hoved- og bidiagnose som similaritetsmål.

C.5 Koblede patientvariable

C.5.1 Alder og køn

Clusterfordelingen med alder koblet med køn er angivet i tabel C.21.

C.5.2 Alder og afdeling

Clusterfordelingen med alder koblet med afdeling er angivet i tabel C.22.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	57.34266%	42.65734%	143
Cluster 2	67.29858%	32.701424%	211

Tabel C.21: Clusterfordeling ved 2 clusters

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	68.200836%	31.799164%	239
Cluster 2	53.04348%	46.95652%	115

Tabel C.22: Clusterfordeling ved 2 clusters

C.5.3 Køn og afdeling

Clusterfordelingen med køn koblet med afdeling er angivet i tabel C.23.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	65.76271%	34.23729%	295
Cluster 2	50.847458%	49.152542%	59

Tabel C.23: Clusterfordeling ved 2 clusters

C.5.4 Alder, køn og afdeling

Clusterfordelingen med koblingen mellem alder, køn og afdeling er angivet i tabel C.24.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	65.76271%	34.23729%	295
Cluster 2	50.847458%	49.152542%	59

Tabel C.24: Clusterfordeling ved 2 clusters

C.5.5 Alder, køn og diagnose (ukategoriseret)

Clusterfordelingen med koblingen mellem alder, køn og diagnose (ukategoriseret) er angivet i tabel C.25.

C.5.6 Alder, køn og diagnose (kategoriseret)

Clusterfordelingen med koblingen mellem alder, køn og diagnose (kategoriseret) er angivet i tabel C.26.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	61.780106%	38.219894%	191
Cluster 2	65.03068%	34.969326%	163

Tabel C.25: Clusterfordeling ved 2 clusters

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	61.70213%	38.29787%	329
Cluster 2	84.0%	16.0%	25

Tabel C.26: Clusterfordeling ved 2 clusters

C.5.7 Alder, afdeling og diagnose (ukategoriseret)

Clusterfordelingen med koblingen mellem alder, afdeling og diagnose (ukategoriseret) er angivet i tabel C.27.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	64.19753%	35.802467%	324
Cluster 2	53.333332%	46.666668%	30

Tabel C.27: Clusterfordeling ved 2 clusters

C.5.8 Alder, afdeling og diagnose (kategoriseret)

Clusterfordelingen med koblingen mellem alder, afdeling og diagnose (kategoriseret) er angivet i tabel C.28.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	63.323784%	36.676216%	349
Cluster 2	60.0%	40.0%	5

Tabel C.28: Clusterfordeling ved 2 clusters

C.5.9 Køn, afdeling og diagnose (ukategoriseret)

Clusterfordelingen med koblingen mellem køn, afdeling og diagnose (ukategoriseret) er angivet i tabel C.29.

C.5.10 Køn, afdeling og diagnose (kategoriseret)

Clusterfordelingen med koblingen mellem køn, afdeling og diagnose (kategoriseret) er angivet i tabel C.30.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	65.87838%	34.12162%	296
Cluster 2	50.0%	50.0%	58

Tabel C.29: Clusterfordeling ved 2 clusters

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	63.532764%	36.467236%	351
Cluster 2	33.333332%	66.666664%	3

Tabel C.30: Clusterfordeling ved 2 clusters

C.5.11 Alder, køn, afdeling og diagnose (ukategoriseret)

Clusterfordelingen med koblingen mellem alder, køn, afdeling og diagnose (ukategoriseret) er angivet i tabel C.31.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	65.87838%	34.12162%	296
Cluster 2	50.0%	50.0%	58

Tabel C.31: Clusterfordeling ved 2 clusters

C.5.12 Alder, køn, afdeling og diagnose (kategoriseret)

Clusterfordelingen med koblingen mellem alder, køn, afdeling og diagnose (kategoriseret) er angivet i tabel C.32.

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	63.532764%	36.467236%	351
Cluster 2	33.333332%	66.666664%	3

Tabel C.32: Clusterfordeling ved 2 clusters

C.6 Kobling krydsede variable

C.6.1 Alder, afdeling og idf-vægtede termer

Clusterfordelingen med koblingen mellem alder, afdeling og idf-vægtede termer er angivet i tabel C.33.

C.6.2 Alder, afdeling og idf-vægtede sidste 10% af samtlige notater uden lemmatisering

Cluster	udskrivninger	behandlingsskift	i alt
Cluster 1	68.04979%	31.950207%	241
Cluster 2	53.097343%	46.902657%	113

Tabel C.33: Clusterfordeling ved 2 clusters