UNIVERSITETET I OSLO
Institutt for informatikk

Would you like a second opinion? Bruk av beslutningsstøtte i helsesektoren

Anders Johan Hem Halling

Masteroppgave,
IT-SLP
30 studiepoeng

23. mai 2007

Sammendrag

Forskning i beslutnings- og bedømmingspsykologi har vist at statistisk prediksjon har potensiale til å forbedre prediktiv beslutningstaking på mange områder. Dette har foreløpig fått liten innvirkning på praktisk beslutningstaking, for eksempel i helsesektoren, på tross av at forskningen nå i 60 har vist entydige fordeler. For å forklare liten bruk må man se på faktorer som menneskelige avvik fra rasjonalitet, strukturelle faktorer ved utformingen av beslutningsstøtte, mellommenneskelige forhold og forholdet mellom menneske og maskin. Forsøk på å påvirke slike faktorer for å øke bruk av beslutningsstøtte har vist seg vanskelige, men forskningen på området er foreløpig begrenset. Ytterligere forskning anbefales spesielt å fokusere på pasientutfall ved bruk av beslutningsstøtte i helsesektoren.

Innhold

Forord

Denne oppgaven er en såkalt "kort" oppgave, noe som innebærer at problemstillingen er gitt av veileder. Problemstillingens ordlyd er som følger:

> "Funn i beslutningspsykologi skulle tilsi en langt større bruk av beslutningsstøtte i mange typer profesjonsutøvelse enn hva som faktisk er tilfelle: Hva kan grunnene være til at beslutningsstøtte ikke brukes mer og hvordan få profesjonsutøvere til å ta slik støtte i bruk? Du må gjerne avgrense diskusjonen til et bestemt fagområde og/eller (en) bestemt(e) form(er) for beslutningsstøtte."

Veileder er Geir Kirkebøen ved Institutt for Psykologi, Universitetet i Oslo.

Oppgaven er strukturert som følger: I del I går jeg gjennom forskningen de siste 60 år, og ser på undersøkelser vedrørende bruk av beslutningsstøtte. Kapittel 2 gjennomgår en del større metastudier som demonstrerer potensialet ved statistisk prediksjon. Kapittel 3 gjennomgår forsøk på å iversette tiltak med bakgrunn i statistisk prediksjon. I del II diskuterer jeg i kapittel 4 om forskningen *bør* ha konsekvenser for klinisk profesjonsutøvelse, i kapittel 5 mulige årsaker til underbruk av beslutningsstøtte og i kapittel 6 mulige tiltak for å øke bruken.

Del I

Introduksjon og Bakgrunn

Kapittel 1

Introduksjon

> During a visit to a mental institution, a visitor asked the Director what the criterion was which defined whether or not a patient should be institutionalized. "Well", said the Director, "we fill up a bathtub, then we offer a teaspoon, a teacup and a bucket to the patient and ask him or her to empty the bathtub." "Oh, I understand", said the visitor. "A normal person would use the bucket because it's bigger than the spoon or the teacup." "No." said the Director, "A normal person would pull the plug. Would you like a bed near the window?"

I litteraturen om beslutning og bedømmingspsykologi er det få debatter som har skapt like store bølger som debatten om aktuarisk (det vil si frekventistisk) versus "klinisk" beslutningstaking. Debatten startet for alvor i 1954 med Meehls bok *Clinical versus statistical prediction*. Denne boka inneholdt blant annet den første metastudien på området, og konkluderte med at statistiske prediksjonsregeler er bedre til å forutsi probabilistiske utfall, herunder menneskelig adferd, enn menneskelige beslutningstakere. Siden har debatten rast, og stadig mer forskning har kommet til. Meehls konklusjoner er validert gang på gang og en skulle derfor tro at statistisk prediksjon og andre former for beslutningsstøtte i dag ville være i utstrakt bruk på en mengde områder. Det har imidlertid vært omfattende og mangfoldig motstand mot innføringen og bruken av beslutningsstøtteverktøy. Denne oppgaven ser på mulige årsaker til og botemidler for denne motstanden, og retter seg spesielt mot helsesektoren (både somatikken og psykiatrien).

For å illustrere hva jeg mener med beslutningsstøttesystemer i klinisk praksis kan vi tenke oss en situasjon der en pasient kommer inn til sin kommunale legevakt med et sett symptomer. Symptomene og legens diagnose vil stort sett

alltid legges inn i et elektronisk pasientsystem. En rolle for beslutningsstøtte kan da være at et sett med statistiske prediksjonsregler på bakgrunn av symptomer og diagnose presenterer en liste med differensialdiagnoser basert på statistikk over hvor vanlige de er. For eksempel "10% av pasientene med denne initialdiagnosen viser seg etter utredning å ha "diagnose B" i stedet." Legen kan da kanskje rekvirere en test for utelukke diagnose B, og får dermed en individuell "huskeliste" for hver pasient basert på objektiv statistikk.

Pasienter presenterer ofte symptomer som kan være konsistente med flere sykdommer. Hver av disse sykdommene har flere ulike behandlingsalternativer med ulik pris, ulike bivirkninger, og ulik prognose. Hvordan kan legen best mulig unngå å feildiagnostisere og/eller feilbehandle denne pasienten, og samtidig ta hensyn til ressursbruk? De siste 60 år har det pågått en debatt i forskningslitteraturen i beslutnings og bedømmingspsykologi angående hvordan vi best kan avgjøre den typen problemer som vi her stilles overfor. Denne kontroversen har stått mellom to fronter, de som mener klinisk skjønnsmessig vurdering er best, og de som mener aktuarisk, statistisk prediksjon er best. Den aktuariske siden har presentert overveldende evidens for at statistisk prediksjon er like bra som eller bedre enn klinisk vurdering, men dette har sett svært liten andvendelse. Denne oppgaven ser på litteraturen på emnet de siste 50 år, og vil så se på mulige årsaker til den paradoksalt lille utbredelsen av beslutningsstøtte i klinisk praksis.

Det finnes flere typer beslutningsstøtte, fra regler og prosedyrer, via statistiske prediksjonsregler (SPR) til enorme dataprogrammer som kalles ekspertsystemer. Av praktiske årsaker har jeg valgt å konsentrere meg i hovedsak om statistiske prediksjonsregler, og det er primært dette som menes med ordet "beslutningsstøtte" i denne oppgaven. Statistiske pediksjonsregler tar ofte form av reglresjonsmodeller, det vil si at man utfører en regresjonsanalyse på statistiske data for å avdekke hvilke variabler som er assosiert med utfallet og i hvilken grad de forskjellige variablene bør vektes.

Et eksempel er en regel for å anslå sannynsligheten for at en person er positiv til aktiv dødshjelp hentet fra Sieck & Arkes, 2005. Regelen er basert på demografiske data, nærmere bestemt kvanifiserte opplysninger om en persons alder (A), partitilhørighet (P), alkoholforbruk (C), religiøsitet (R) og holdning til sex før ekteskapet (S). Ved regresjonsanalyse har man kommet fram til formelen $Y = 0.31 - 0.02A - 0.26P - 0.37C + 0.21R + 0.63S + 0.30(P * C) - 0.27(C * R)$ Denne formelen viste seg å ha en prediktiv treffsikkerhet på 77%, som var mer enn noen av faktorene enkeltvis.

Når det er snakk om beslutningsstøttesystemer mener man som regel at statistiske prediksjonsregler er integrert i et dataprogram som kan være enkeltstående eller som kan være en del av andre datasystemer som er i daglig bruk der man ønsker å innføre beslutningsstøtten. Utbredelsen av slik beslutningsstøtte er langt mindre enn forskningen skulle tilsi, og litteratur som beskriver forsøk på å innføre dette har vært lite oppløftende lesning (se for eksempel Arkes, 2007 eller Arkes, 2003). Jeg vil derfor se litt på mulige årsaker til disse funnene, psykologisk og organisatorisk, og også vurdere eventuelle botemidler som kan være effektive for å lette innføringen av beslutningsstøtte. For ikke å favne alt for vidt har jeg valgt å konsentrere meg om beslutninger og beslutningsstøtte relatert til helsesektoren.

Beslutningsstøttesystemer kan bidra på flere områder i helsesektoren. De kan assisstere ved diagnostisering, valg av behandling, overvåking (av for eksempel prøvesvar eller medikamentinteraksjoner), eller ved prediksjon av fremtidig adferd (hvor sannsynlig er det at en voldelig psyiatrisk pasient med denne diagnosen vil utøve vold igjen hvis han skrives ut nå?). Beslutningspsykologisk forskning har vist at denne typen beslutningsstøtte kan redusere feildiagnostisering og feilbehandling med en liten, men konsistent, andel på en rekke områder (Dawes et al., 1989; Hunt et al., 1998; Grove et al., 2000; Ægisdottir et al., 2006).

Kapittel 2

Statistisk versus klinisk predkisjon

Ordet "klinisk" som alternativet til "statistisk" prediksjon er ikke nødvendigvis heldig i forhold til å overtale klinikere til å ta statistiske prediksjonsregler mer i bruk. Jeg har allikevel valgt å benytte dette ordet, da det er i utstrakt bruk gjennom hele litteraturen. Dette er på ingen måte ment å gi inntrykk av at klinikere generelt ikke fatter beslutninger til beste for sine pasienter.

Når jeg snakker om prediksjon mener jeg nærmere bestemt prediksjon av probabilistiske eller usikre utfall. I helsevesenet kan denne formen for beslutninger for eksempel ta form av "Hvor stor er sjansen for at symptom A betyr at pasienten lider av sykdom B", eller "Hvor stor er sjansen for at behandling A vil lykkes? (evt. versus behandling B?)" Slike spørsmål har i virkeligheten ett enkelt riktig svar. Pasienter har enten sykdom B eller ikke. Dette betyr imidlertid ikke at problemet kan behandles som et eksploratorisk problem der vi går gjennom alle mulige alternativer. Antallet mulige alternativer er gjerne alt for stort, det eksisterer kanskje ikke definitive tester som kan avdekke alle disse tilstandene, og det vil ofte være alt for ressurskrevende å utrede alle pasienter så grundig. I praksis anses diagnostiske problemer derfor ofte som probabilistiske. Svaret på et probabilistisk problem kan dermed sjelden formuleres som ja/nei, men må i stedet formuleres som sannsynligheter som "sannsynligheten for at pasienten har sykdom B er 0.85 (85%)". For å komme frem til dette svaret må vi ha noen data om pasienten, og en måte å kombinere disse dataene på. Det er kun den kombinatoriske fasen som er tema i denne oppgaven.

Det er to hovedmåter som brukes for å kombinere data om pasienten, klinisk og aktuarisk/statistisk metode. Med klinisk prediksjon menes at data om

pasienten kombineres til en sannsynlighetsvurdering inne i hodet på en menneskelig beslutningstaker basert på skjønn. Med aktuarisk prediksjon menes at data om pasienten kombineres etter eksplisitte kriterier, gjerne i en matematisk formel. Et eksempel kan være regler som Glascow Coma Scale (GCS) (Teasdale and Bennet, 1974) som er et verktøy for å vurdere bevisstheten hos pasienter. GCS baserer seg på tre kriterier: Vudering av pasientens beste motoriske nivå, på en skala fra 6-1 ("Normale bevegelser", "Målretter avverge ved smerte", "Målrettet tilbaketrekking ved smerte", "Fleksjon ved smerte", "Ekstensjon ved smerte", "Ingen motorisk respons") vurdering av pasientens beste verbale nivå, fra 5-1, ("Normal samtale", "Desorientert samtale", "Tilfelige ord", "Lyder", "Ingen verbal respons" og vurdering av pasientens beste øyerespons på en skala fra 4-1 ("Åpner øynene spontant", "Åpner øynene ved tiltale", "Åpner øynene ved smerte" og "Ingen øyerespons"). I bruk oppgir man så vurderingen på hvert enkelt kriterie, sammen med summen av alle tre. For eksempel "øyne:4, verbalt:4, motorisk:6 = GCS:14" Hvis man også har statistikk om hvordan det tidligere har gått med pasienter basert på GCS ved innkomst til sykehus, og GCS viser seg å ha sann prediktiv kraft ut fra denne statistikken, kan man så lage en SPR som kombinerer GCS-data og statistiske data for å predikere pasientens prognose. Slike regler kan for eksempel se slik ut: "En pasient med hodeskade som har GCS på under 8 ved innkomst har 40% til 50% sjanse for å overleve" "En pasient med hodeskade som har GCS over 8 ved innkomst har over 80% sjanse for overlevelse". Hvis to slike pasienter kommer samtidig til samme sykehus kan en slik regel hjelpe legene i akuttmottaket med prioritering av hvem som må behandles først.

Det har lenge vært kjent at selv enkle statistiske prediksjonsregler overgår menneskelige vurderinger på en rekke felt der slike probabilistiske beslutninger må fattes (se for eksempel Dawes, Faust og Meehl, 1989). På tross av dette har denne typen beslutningsstøtte ikke blitt brukt i noen særlig grad "ute i verden" (Arkes et al., 2007). I denne seksjonen gjennomgås forskningen på dette feltet, slik den står i dag. Vi skal se på evidensen for, og de antatte årsakene til, at statistiske prediksjonsregler konsistent tangerer eller overgår menneskelig klinisk prediksjon, og se litt på problemer knyttet til bruken av beslutningsstøtte generelt.

2.1 Begynnelsen - Meehl skriver en forstyrrende liten bok

Paul Meehls bok *Clinical versus statistical prediction* fra 1954, inneholdt et kapittel der han gikk gjennom de studier som da fantes (20), som sammenlignet menneskelig skjønn med statistisk prediksjon. Hans konklusjon var at statistisk prediksjon var gjennomgående like bra som, eller bedre enn, klinisk prediksjon. Denne forskningen fikk imidlertid liten oppmerksomhet utenfor forskningskretser, og Meehl sier også at det var et "vanntett skott" mellom psykologisk forskning og terapeutisk praksis (Meehl, 1986). Innvendinger mot statistisk prediksjon ble også reist, og Meehl (1986) nevner for eksempel: studiene som var inkludert var vektet til fordel for aktuariske metoder blant annet ved at metodene ikke hadde adgang til lik informasjon, at informasjonen som var tilgjengelig var av en slik karakter at den ga aktuarisk metode en fordel, at det var valgt ut kunstige eller lite relevante prediktive oppgaver, eller at klinikere som deltok ikke hadde nødvendig kompetanse til å gjøre den type prediksjoner som var omfattet. Det ble også hevdet at aktuarisk prediksjon ikke er fleksibel nok til å ta hensyn til eksepsjonelle tilfeller som et menneske ville sett øyeblikkelig. Om vi for eksempel har en SPR som predikerer seilingstiden til danskebåten på strekningen Oslo-Fredrikshavn-Oslo, som tar hensyn til for eksempel vær og bølgehøyde, vil denne prediksjonen kunne bomme grovt ved sjeldne hendelser som mann-over-bord eller brann ombord. Disse hendelsene vil være så sjeldne at de neppe er omfattet i prediksjonsregelen (Og antall mulige hendelser som kan forsinke en danskebåt går mot uendelig, så det vil også være umulig å inkludere dem alle). En menneskelig beslutningstaker vil derimot straks skjønne at en slik hendelse er relevant i forhold til prediksjon av seilingstid. Dette kalles i litteraturen "the broken leg problem", og man skulle tro at dette ga klinisk prediksjon en åpenbar fordel. Det som har vist seg er imidlertid at et menneske ofte finner mange flere avvik enn det er grunnlag for i dataene. Vi har altså en tendens til å tro at unike hendelser har større innvirkning på utfallet enn det som er tilfellet. Der aktuarisk prediksjon overser sjeldne og unike hendelsers innflytelse på utfall, har klinisk prediksjon en tendens til å overvudere betydningen av disse unike hendelsene (Dawes, Faust & Meehl, 1989, s. 1670).

For å imøtegå kritikerne presenterte Meehl (1954, i Meehl, 1986) to kriterier han mente må tilfredsstilles før en sammenlignende studie kunne sies å være rettferdig: For det første må begge metodene ha tilgang til, eller være basert på, det samme datasettet. Begge metodene trenger dog ikke å bruke det fullstedige datasettet. Det typiske er at den aktuariske metoden er ba-

sert på en delmengde av dataene som er tilgjengelig for den menneskelige beslutningstakeren. For det andre må man unngå forhold som kan føre til kunstig høye resultater for statistisk prediksjon. Dette vil for eksempel si at en prediksjonsregel som er basert på ett sett med kjente utfall i en kjent populasjon, må kryssvalideres mot en annen representativ populasjon for å sikre at prediksjonsregelen er generaliserbar og ikke representerer tilfeldige sammenhenger i det opprinnelige datasettet.

2.2 Utålmodighet - Hypotesen styrkes

Etterhvert som det kom mer forskning, som overveldende støttet Meehls konklusjoner, så man fortsatt ingen utstrakt anvendelse av disse konklusjonene. Mot slutten av 80-tallet endret litteraturen karakter fra kun å konstatere at statistisk prediksjon fungerer, til også å søke å overtale klinikere til å ta dette i bruk. En sentral artikkel her er Dawes, Faust og Meehls Science-artikkel *Clincal versus actuarial judgement* fra 1989. De argumenterer som tidligere nevnt for å skille prediktive oppgaver i to distinkte faser. En datainnsamlingsfase, og en kombineringsfase. De fremholder, som Meehl har ment hele tiden, at menneskets evne til å oppfatte nyanser, mønstre og tegn, i medisinen det som kalles "klinisk blikk", er uovertruffen. Menneskets rolle som datainnsamler er suveren og utfordres på ingen måte. På den annen side er menneskets evne til å kombinere en mengde variabler på en konsistent måte heller laber, noe som illustreres av et sitat fra Meehl, 1986:

> "Surely we all know that the human brain is poor at weighting and computing, When you check out at a supermarket, you don't eyeball the heap of purchases and say to the clerk, "Well, it looks like about $17.00 worth to me, what do you think?" The clerk adds it up."
>
> — *Meehl, 1986; s. 372*

Dawes, Faust og Meehl mener derfor at mennesket burde utnytte sine sterke sider, som oppdagelse og gjenkjenning av relevante variabler, og overlate til en SPR å komme frem til hva dataene samlet betyr, for eksempel for et sykdomstilfelles videre forløp. De anså evidensen for at statisktisk prediksjon overgår klinisk som overveldende, og gjentok Meehls konklusjon fra 1986:

> "There is no controversy in social science that shows such a large body of qualitatively diverse studies coming out so uniformly in

> the same direction as this one. When you are pushing 90 investigations, predicting everything from the outcome of football games to the diagnosis of liver disease and when you can hardly come up with half a dozen studies showing even a weak tendency in favour of the clinician, it is time to draw a practical conclusion"
>
> — *Meehl, 1986; ss. 374-375*

2.3 Status presens - 50 år med forskning

Da Meehl skrev sin lille bok i 1954 fant han 20 studier å basere sitt empiriske kapittel på. De metastudiene jeg skal oppsummere her er basert på over 150 studier. Grove et al. (2000) inkluderer 136 studier innen medisin og psykologi, og Ægisdottir et al. (2006) inkluderer 69 studier innen klinisk psykologi (med noe overlapp, Ægisdottir et al. inneholder 28 artikler som ikke er omfattet av Grove et al.).

2.3.1 Grove et al.

Groves store metastudie (Grove & Meehl, 1996; Grove et al., 2000) inkluderte altså 136 studier fra 1920-tallet og frem til 1986. Alle studiene kom fra feltene medisin og psykologi, og alle predikerte "noe om mennesker", som tilstander, forløp eller fremtidig adferd. Av disse studiene konkluderte 63 med klart bedre prediktiv kraft (definert som mer enn 0.1 høyere effektstørrelse) for aktuariske metoder, 65 studier konkluderte med circa like stor prediktiv kraft for begge, og 8 studier konkluderte med klart bedre prediktiv kraft for kliniske metoder. Undersøkelse av disse siste 8 avdekket ingen spesielle domener eller spesielle typer beslutningstakere (for eksempel leger), og hadde generelt veldig lite til felles. Dette er et viktig poeng, da mangel på slike fellestrekk svekker hypotesen om det finnes definerbare fellestrekk ved enkelte domener som predikerer klinisk overlegenhet. Med så få studier til fordel for klinisk prediksjon, og ingen systematikk i disse, forklares disse funnene som dels uttrykk for statistiske artefakter og dels ved det faktum at i 7 av disse 8 studiene hadde den menneskelige beslutningstakeren tilgang til ekstra informasjon som ikke var inkorporert i den aktuariske metoden. Det å ha tilgang til ekstra informasjon var imidlertid generelt ingen fordel for klinisk prediksjon, og hvis denne ekstra informasjonen var data fra et intervju påvirket det nøyaktigheten *negativt*. I disse 7 enkeltsakene kan det dog være at den ekstra informasjonen klinikerne hadde tilgang til hadde sann prediktiv verdi. Andre faktorer som påvirket prediktiv treffsikkerhet var bruk av medisinske data

som laboratorietester og kliniske undersøkelser. Bruk av slike data reduserte forskjellen mellom metodene noe, men endret ikke konklusjonen.

Bruk av psykologiske data, beslutningstakerens utdanning og erfaring, eller tilgang til ekstra informasjon, hadde derimot ingen innvirkning på treffsikkerhet.

Etter grundig analyse av effektstørrelsene fant Grove et al. at den gjennomsnittlige gevinsten ved bruk av aktuarisk metode førte til en økning i prediktiv treffsikkerhet på rundt 13%.

2.3.2 Ægisdottir et al.

Ægisdottir et als studie fra 2006 var konstruert for å svare på spørsmål og innvendinger de mente Grove et al. ikke hadde adressert. Blant disse var innvendinger om at Groves metastudie var for bred, all den tid de hadde inkludert studier fra en rekke forskjellige domener som økonomi, psykologi og medisin, og at dette på ett eller annet vis svekket funnenes kraft i klinisk psykologi. Innvendingen gikk ut noe sånt som at funn på områder som økonomi og klinisk medisin ikke sa noe om relevans for klinisk psykologi, en innvending som ikke tar hensyn til at Grove et al. viste at dette er generelle trekk ved klinisk versus aktuarisk prediksjon. Ægisdottir et al. ønsket derfor å konsentrere sin undersøkelse rundt studier i klinisk psykologi, for å gi sine funn større relevans for utøvende psykologer.

De 69 studiene som ble inkludert var fra 1940-1996, og dekket dermed en del nyere studier som ikke var omfattet av Grove et als arbeid. Studien undersøkte den relative treffsikkerheten til klinisk og statistisk prediksjon med hensyn til intet mindre enn 12 uavhengige variable. Variabler som prediksjonstype (diagnose, prognose, fremdtidig adferd, liggetid på sykehus og så videre (disse diskuteres nærmere nedenfor)), type av statistisk prediksjonsformel (lineære modeller, logiske regler, modeller av klinisk beslutningstaking), informasjonsmengde (var begge strategiene basert på samme data eller hadde en av dem mer informasjon?), tilgjengelighet til objektive baserater, klinikernes tilgang til output fra den statistiske regelen, og så videre.

Målet med studien var å avgjøre hvordan kliniske psykologer (counseling psychologists) optimalt sett burde fatte beslutninger; under hvilke forhold klinisk prediksjon er å foretrekke, under hvilke forhold statistisk prediksjon er å foretrekke, og under hvilke forhold forskjellen mellom de to er så liten

at det ikke spiller noen rolle. Det var slike konkrete anbefalinger de savnet hos Grove et al.

Ægisdottir et als undersøkelse av disse studiene avdekket mye det samme som Grove et al. På grunn av svært konservativ behandling av dataene hadde de en lav terskel for å ekskludere utliggere, og SPR som ikke var kryssvalidert. De endte dermed opp med å inkludere 41 artikler, som ga 48 effektstørrelser (enkelte artikler inneholdt mer enn ett eksperiment) som fordelte seg som følger: 25 (52%) viste klar fordel for statistisk prediksjon, 18 (38%) viste ingen forskjell, mens 5 (10%) viste klar fordel for klinisk prediksjon. Gjennomsnittlig fant de at statistiske metoder kunne øke prediktiv treffsikkerhet fra 47% til 53% (altså gjennomsnittet av prediktiv treffsikkerhet for klinikerne i samtlige studier versus gjennomsnittet for statistiske metoder i samtlige studier), og effekten var statistisk signifikant (med utliggere og ikke-kryssvaliderte SPR inkludert økte forspranget, det vil si at brorparten av de ekskluderte utliggerne gikk i favør av statistisk prediksjon).

Når det gjelder de uavhengige variablene som inngikk i undersøkelsen var det en del interessante funn jeg vil gå nærmere inn på.

Type prediktiv oppgave: Det ble antatt at forskjellen mellom statistiske og kliniske metoder ville variere avhengig av hva slags prediktiv oppgave som ble testet, og dette viste seg å stemme. Statistiske metoder var bedre når det gjaldt å predikere forløp/prognose, forbrytelse/vold og fremtidig akademisk prestasjon. For resten av de prediktive oppgavene (diagnose, liggetid på sykehus, deteksjon av hjerneskade, prediksjon av IQ, personlighetstype, klassifisering av MMPI-profil, prediksjon av selvmordsforsøk og prediksjon av homoseksualitet) var det ingen systematisk forskjell eller det var ikke tilstrekkelige data i de opprinnelige studiene til å analysere dette. Ingen prediktive oppgaver favoriserte klinisk prediksjon.

Datainnsamling: En innvending som var reist mot klinisk prediksjon var at dataene som ble presentert for de menneskelige beslutningstakerne var helt fremmede for dem, og at de derfor ikke kunne yte sitt beste (Holt, 1970 i Ærgirsdottir et al., 2006). Ægisdottir et al. sammenlignet derfor studier der klinikerne ble testet i data fra samme klinikk som de praktiserte i til daglig, med studier der klinikere ble testet på (for dem) ukjente data. Hypotesen var at klinikere bør vite hvordan de skal integrere data som stammer fra deres egen klinikk/institusjon. De fant en sammenheng, men ikke helt som de hadde trodd. Det viste seg at det å

kjenne dataene fra før var assosiert med *dårligere* treffsikkerhet enn om man hadde blitt presentert med data fra en annen klinikk/institusjon. "Thus—and most unexpected—existing studies indicate that clinicians seem to be more accurate when they are working with less familiar or novel information." (Ægisdottir et al., s. 363)

Type statistisk formel: Alle typer statistisk prediksjon gjorde det bedre enn klinkerne. Det interessante var dog at for metoden "logisk konstruerte regler" var forskjellen ikke signifikant. Det var kun de rent statistiske metodene som viste signifikant forbedring. Et eksempel på logisk konstruerte regler er Goldbergs regel for å skille mellom nevroser og psykoser. Denne regelen baserer seg på en personlighetstest (MMPI), der man tar summen av tre skalaer og trekker fra summen av to andre. Er sluttsummen mindre enn 45 klassifiseres pasienten som nevrotisk, er den over 45 klassifiseres pasienten som psykotisk (pasienten er forhåpentligvis allerede klassifisert som enten nevrotisk eller psykotisk, ellers ville dette resultert i en smule overdiagnostidering...).

Informasjonsmengde: Holt (1970 i Ægisdottir et al., 2006) har argumentert med at klinikerne trenger mer og annerledes informasjon enn det rent statistiske for virkelig å vise hva de kan. Han mente at meningsfull kvalitativ informasjon var et must. I likhet med Dawes et al. (1989) og Grove et al. (2000) mente Ægisdottir et al. (2006) at dette ikke ville ha noen effekt. Det viste seg at tilgang til mer informasjon, i Ægisdottir et als tilfelle stort sett intervjuer, var assosiert med negativ prediktiv kraft.

Baserateinformasjon: Det var ingen forskjell i treffsikkerhet basert på hvorvidt klinikerne hadde tilgang til baserater eller ikke. Dette var som antatt, da det tidligere er demonstrert at baserateinformasjon ikke vektlegges tilstrekkelig (Bar-Hillell, 1980). Å gi klinikere baserateinformasjon alene er derfor ikke noen god debiasingteknikk.

Tilgang til SPR: Det var ingen forskjell i treffsikkerhet avhengig av om klinikerne hadde tilgang til resultatene fra den statistiske metoden. Som tidligere nevnt kan dette være fordi klinikerne da aktivt lette etter "brukne bein", og dermed fant flere slike enn det er grunnlag for. Det er også dokumentert at selv der beslutningsstøtte er tilgjengelig konsulteres den sjelden (Arkes et al., 2007; Sieck and Arkes, 2005)

Klinisk ekspertise: I 7 av de undersøkte studiene var de kliniske beslutningstakerne ansett å være eksperter på sine felt. For disse 7 studiene

var det ingen signifikant forskjell på klinisk og statistisk prediksjon (effektstørrelsen viste en fordel for statistisk prediksjon også her, men konfidensintervallet inkluderte 0). Sann ekspertise (som vurdert av andre) virker dermed å være en faktor som har sann prediktiv kraft, men her må en sannsynligvis legge til grunn en smal definisjon av ekspertise, da Grove et al. (2000) jo ikke fant noen sammenheng mellom utdanning/erfaring og treffsikkerhet. Ægisdottir et al. har desverre ikke oppgitt hvilke prediktive oppgaver disse studiene omhandlet, eller noen definisjon av hva de legger i ordet ekspertise.

Totalt sett fant Ægisdottir et al. at gjennomsnittlig forskjell i treffsikkerhet var den samme som hos Grove et al., 13% i favør av statistisk prediksjon. To metastudier, utført med mindre enn 10 års mellomrom, med ulike studieutvalg (det var 28 studier i Ægisdottir et al. (2006) som ikke var omfattet av Grove et al. (2000)) ender altså opp med identiske resultater. Disse funnene kan derfor anses som konsistente og robuste og gjeldende i en lang rekke domener. Grove et al. (2000) går så langt som til å hevde at enhver ny metastudie som utføres vil gi tilsvarende resultater:

> "The trend in our data is so strong that we conjecture the following: There is no selection of studies, based on anything except study outcome itself, that will yield a conclusion directly contrary to ours."
>
> — *Grove et al., 2000; s. 26*

Ægisdottir et al. (2006) ønsket som tidligere nevnt å presentere konkrete anbefalinger til praktiserende psykologer. Anbefalingene de gir er:

(1) Generelt bruker de sine egne og Grove et als konklusjoner til å anbefale bruk av SPR overalt der det er praktisk mulig ("when feasible"), og spesielt hvis treffsikkerhet er viktig, og konsekvenser av feil er store.

(2) Ikke alle typer prediksjonsregler fungerer. Logisk konstruerte regler som Goldbergs MMPI-regler for diagnostisering av nevroser og psykoser (Goldberg, 1965 i Ægisdottir et al., 2006) var ikke bedre enn klinisk skjønn (regelen fungerte for Goldbergs populasjon, men generaliserte dårlig til andre populasjoner). Kun "ekte" statistiske modeller viste forbedring.

(3) Praktiserende psykologer burde aktivt søke informasjon om eksisterende SPR. Som vist nedenfor kan det å ignorere slike hjelpemidler der de eksisterer betegnes som uetisk (Dawes, 2002, 2005).

(4) Det følger også at de må tillegge disse hjelpemidlene stor vekt, og faktisk benytte dem.

(5) Baserateinformasjon er vist å kunne ha effekt dersom klinikerne blir trent i å benytte denne informasjonen (Spengler et al., 1995 i Ægisdottir et al., 2006). Slik informasjon er ofte tilgjengelig for praktiserende psykologer gjennom DSM-IV-TR, men det trengs mer forskning for å finne ut hvordan man best kan trenes opp til å benytte den informasjonen riktig.

(6) Praktiserende psykologer bør bli oppmerksomme på begrensningene som er vist i forhold til klinisk prediksjon, selv der de jobber med kjent informasjon i en kjent kontekst. De bør med andre ord gjøre seg kjent med forskningen på debiasing. Som foreslått av blant annet Klein (1999) kan det virke som at en vanlig menneskelig strategi er å konstruere en hypotese for så å søke bekreftende informasjon, i stedet for den falsifiserende metoden hypotetisk-deduktiv metode inviterer til.

(7) Til slutt ramser de opp noen felt der de mener forskningen viser at praktiserende psykologer trygt kan bruke klinisk prediksjon, med andre ord der det ikke er vist noen (signifikante) fordeler med statistisk prediksjon. Dette er felter som prediksjon av antatt liggetid/behandlingstid, diagnostisering og vurdering av personlighetstrekk eller hjerneskade ut fra testresultater. Det kommenteres at slike prediktive oppgaver karakteriseres av en relativt god feedback-syklus, som tillater læring om sanne prediktive variabler over tid (mer om feedback i 2.4.3).

2.4 Mulige årsaker til aktuarisk overlegenhet

Vi har nå sett at det er en massiv overvekt av forskningsresultater som underbygger hypotesen om at aktuarisk prediksjon som oftest er like bra eller bedre enn klinisk prediksjon. Årsakene til disse funnene er derimot ikke like godt dokumentert. Det er lansert en del teorier, og jeg vil nå ta for meg et par av disse.

2.4.1 Svakheter ved klinisk skjønn

Burde disse funnene egentlig overraske oss? Det har lenge vært kjent at menneskelige beslutningstakere avviker fra perfekt rasjonalitet på en rekke punkter (se for eksempel Plous, 1993). En retning i beslutnings og bedømmingspsykologi er faktisk dedikert nettopp til å utforske forskjellige avvik fra ra-

sjonalitet, under hvilke forhold de oppstår, og hva som eventuelt kan gjøres for å motvirke dem. Denne retningen, kalt "heuristics and biases", har dokumentert en rekke slike skjevheter (biases) som kan påvirke nøyaktigheten til helserelaterte beslutningstakere (Plous, 1993). Eksempler kan for eksempel være overkonfidens. Overkonfidens vil si at man anser sin egen prediksjon for å være mer nøyaktig enn den faktisk er. Et eksempel er Ukrainas minister for energi og elektrifisering, som i 1986 uttalte seg om atomsikkerhet:

> "The odds of a meltdown are 1 in 10.000 years."
>
> — *Vitali Skylarov i Plous, 1993 s. 217*

Tre måneder senere eksploderte reaktor 4 i Tsjernobyl, og undersøkelser avdekket en rekke risikofaktorer ved denne typen reaktorer som ikke var tatt hensyn til (Stang, 1996). En kan trygt si at sovjeterne var overkonfidente med tanke på sin egen risikoanalyse. Overkonfidens kan lett føre til at man anser et problem for løst (fordi man jo er så sikker på at man har funnet det riktige svaret) og dermed slutter å lete etter alternative forklaringer. Overkonfidens kan også forklare hvorfor beslutningsstøtte blir lite brukt, og diskuteres grundigere i 5.7.3.

Skjevheter og avvik fra rasjonalitet kan føre til at klinisk prediksjon blir dårligere enn den kunne vært, og dårligere enn man skulle tro om man ikke kjenner til litteraturen om "heuristics and biases". Når man så avviker fra rasjonalitet, så gjøres det heller ikke på en konsistent måte. Tilgjengelighetsheuristikken (Plous, 1993) kan blant annet innebære at hendelser som er lett tilgjengelige, det vil si lette å huske eller komme på, tillegges uforholdsmessig mye vekt, og dette kan bidra til skjev eller inkonsistemt vekting av variabler. La oss se for oss følgende situasjon: En lege har nylig fått inn en pasient med litt diffuse symptomer. Etter lang tids utredning og mange tester finner han ut at pasienten hadde en svært sjelden sykdom. Dette tilfellet var så spesielt at legen husker dette godt i lang tid. Neste gang legen da får en pasient med liknende symptomer vil han da lett kunne overvurdere sannsynligheten for at også denne pasienten har den (svært sjeldne) sykdommen, fordi minnet om den forrige er så lett *tilgjengelig* i hukommelsen. Pasienter kan dermed behandles forskjellig avhengig av hvilke typer saker legen har sett den siste tiden.

Skjevheter og avvik fra rasjonalitet på denne måten kan føre til at klinisk prediksjon blir dårligere enn den kunne vært, og dårligere enn man skulle tro om man ikke kjenner til heuristics and biases-litteraturen. Et av de

store problemene slike skjevheter fører til er nedsatt reliabilitet for klinisk prediksjon.

2.4.2 Reliabilitet

Den årsaken til aktuarisk overlegenhet man oftest hører foreslått, og som på meg virker svært rimelig, går på reliabilitet. En aktuarisk metode vil, per definisjon, alltid gi samme resultater gitt samme input. De har med andre ord perfekt test-retest reliabilitet. Gitt god opplæring og god utforming av SPR i forhold til brukervennlighet bør de også ha god interrater-reliabilitet. For menneskelige beslutningstakere er dette derimot langt i fra tilfellet. Grunnet forekomsten av skjevheter som nevnt i forrige avsnitt kan en psykolog som sitter med to forskjellige pasienter som i bunn og grunn presenterer akkurat de samme symptomene godt komme fram til forskjellige konklusjoner. Interrater-reliabiliteten til to forskjellige psykologer, som til og med kan tilhøre forskjellige "skoler" i psykologien, er som regel også et stykke unna det man kan få til med aktuariske metoder.

Dette poenget blir svært viktig i forhold til å forklare hvorfor aktuarisk prediksjon kommer bedre ut. Validiteten av en test eller prediksjon kan som kjent ikke oversige reliabiliteten. Når aktuarisk prediksjon har tilnærmet perfekt reliabilitet, og klinisk prediksjon ikke har det, vil aktuarisk prediksjon ha en stor fordel, på tross av de åpenbare svakhetene, slik som "broken-leg" problematikken.

2.4.3 Feedback

Det er demonstrert at mennesker generelt er overkonfidente når det gjelder sine egne evner og ferdigheter, det vil si at vi vurderer oss selv som bedre enn vi faktisk er (se 5.7.3 for mer om dette). Jevnlig feedback over lang tid har vist seg til en viss grad å kunne motvirke overkonfidens (Plous, 1993), selv om fenomenet er svært robust. Uten feedback er man derimot nesten garantert overkonfidens, og det er et stort problem på svært mange fagområder at man ikke får feedback, eller man får feedback av varierende kvalitet. Spesielt i helsesektoren, der man også er underlagt taushetsplikt og personvernhensyn, er det ofte vanskelig å tilegne seg informasjon om de utfall man tidligere predikerte uten å gjennomføre dette som særskilte forskningsprosjekter. Menneskelige beslutningstakere i slike domener er derfor ofte dårlig *kalibrert*, det vil si at de ikke har kunnskap om sine egne prediktive ferdigheter (Sieck and Arkes, 2005). Som vi så i Ægisdottir et al. (2006) var tilgang til ekstra informasjon assosiert med dårligere treffsikkerhet. En mulig årsak til dette er da

at klinikerne "vanner ut" sin interne utregning ved å blande variabler med sann prediktiv kraft med "støy" fra de ekstra variablene de har tilgang til. Det var også et fellestrekk ved de typene prediktive oppgaver der forskjellen var liten (ikke signifikant) at de hadde en relativt stabil og god feedback-syklus. Statistisk prediksjon har ikke slike problemer da SPR er utarbeidet nettopp med grunnlag i hvilke variabler som har sann prediktiv kraft, gjerne via en regresjonsanalyse, og hver variabel som skal legges til regelen må demonstrere inkrementell validitet (det vil si at den ikke bare skal ha sann prediktiv kraft, men unik prediktiv kraft som ikke allerede er dekket av en annen variabel).

Kapittel 3

Bruk av beslutningsstøtte

Som vist over er det massiv evidens i forskningen for at statistisk prediksjon kan føre til en moderat, men konsistent, bedring av prediktiv treffsikkerhet på en rekke helserelaterte områder. Disse funnene er ikke nye, og en skulle derfor tro at de hadde hatt mer enn rikelig tid til å bli tatt i bruk. Det har imidlertid vist seg at dette er langt fra tilfellet (Kaplan, 2001; i Arkes, 2007). En del eksempler har også vist at forsøk på å implementere til dels svært enkel beslutningsstøtte har blitt møtt med aggressiv motstand (Arkes, 2003).

Når det gjelder bruk av beslutningsstøtte der den er tilgjengelig har dette også vist seg å være vanskelig. En ny artikkel fra Arkes (2007) oppsummerer problemet slik:

> "Reviews by Kaplan and Hunt and colleagues confirm that although many studies verify the superiority of DSSs [decision support systems] in the diagnostic process, some studies do not. However, there is unanimity with regard to 1 characteristic of DSSs: they are grossly underused. To cite 1 example of underutilization, the acute ischemic heart disease predictive instrument put in place by Corey and Merenstein reduced the false-positive diagnosis rate from 71% to 0%. Following the use of the aid in randomized controlled trials, physicians were free to use the aid or not. Utilization during this latter phase was only 2.8%. Other examples of underutilization abound"
>
> — *Arkes, 2007 s. 190*

3.1 Intens motstand mot innføring av enkle statistiske regler

Et interessant funn er Hal Arkes erfaringer da han forsøkte å innføre noen enkle endringer i metodene som blir benyttet for å vurdere støtte til forskningsprosjekter ved USAs National Institutes of Health, og National Science Foundation. Hal Arkes har forsket på risikopersepsjon og beslutningspsykologi rettet mot helsesektoren i en årrekke, og har publisert en rekke artikler på området. Da han satt i en komite som skulle vurdere søknader om forskningsstøtte anså han det som en gylden mulighet til å se om han kunne anvende funnene fra forskningen i praksis.

3.1.1 National Science Foundation

National Science Foundation (NSF), så vidt jeg kan se motstykket til Norges Forskningsråd, hadde fått kritikk av Government Accounting Office (Riksrevisjonen) for å bruke andre kriterier i vurderingen av søknader om forskningsmidler enn det de selv oppga. Arkes satt i en av komiteene i NSF, og foreslo derfor tiltak basert på forskningen i bedømmings og beslutningspsykologi for å bøte på dette.

Kritikken gikk ut på at NSF brukte andre kriterier i sine vurderinger i tillegg til eller i stedet for de offisielle. ("unwritten or informal criteria were used", Arkes, 2003; s. 1) Dette var naturligvis urettferdig for forskere som ikke kjente systemet fra innsiden. De offisielle kriteriene var: teoretisk grunnlag for forskningen det søktes om støtte til, kvaliteten på søkerens utdanning, nytten av det foreslåtte prosjektet og den foreslåtte metoden.

Hans konkrete forslag var tredelt. Han foreslo at for hvert medlem av tildelingskomiteen skulle man normalisere rangeringen de gav forskjellige søknader de hadde vurdert. Dette ville synliggjøre forskjellen på "strenge" og "snille" komitemedlemmer (medlem A gir prosjekter en gjennomsnittlig skore på 2.5, medlem B gir en gjennomsnittlig skore på 3) og forskjellene ville da kunne korrigeres for, noe som ville føre til bedre interrater-reliabilitet. For det andre foreslo han å undersøke dataene for å finne terskler som kunne identifisere alle søknader som garantert ville bli avslått eller garantert godkjent, så man ikke trengte å bruke masse tid på å diskutere saker med "kjent" utfall (for eksempel at alle saker med gjennomsnittsskore under 1.5 trygt kan avvises uten diskusjon). For det tredje foreslo han at komitemedlemmene skulle rangere søknadene på de 4 eksplisitte kriteriene de allerede hadde, og ikke ved

å rangere søknaden som helhet. Helhetsvurderingen kunne da enkelt regnes ut ved å ta gjennomsnittet av de 4 eksplisitte vurderingene. De burde altså vurdere prosjektets nytte, søkerens utdanning, prosjektets teoretiske grunnlag og metode, og så bare summere eller regne ut gjennomsnittet, i stedet for å vurdere hele prosjektet, da man risikerer at medlem A og B da vekter for eksempel "metode" i forskjellig grad.

Ingen av rådene ble tatt til følge, og det ble i stedet vedtatt at komitemedlemene selv skulle vurdere relevansen til hvert enkelt kriterie i forhold til hver enkelt søknad. Arkes kommenterer at dette nærmest vil garantere at forskjellige komitemedlemmer legger vekt på forskjellige kriterier, og dermed føre til dårlig interrater-reliabilitet.

3.1.2 National Institutes of Health

National Institutes of Health (NIH) hadde fått samme type kritikk som NSF, og Arkes satt som medlem i komiteen som skulle vurdere endringer. Hans konkrete forslag var også her tredelt. NIH evaluerte søknader på en skala fra 1-150. Det er vist at slike vurderingsskalaer når sin beste reliabilitet ved rundt 7 poeng (Landy & Farr, 1980; i Arkes, 2003), og Arkes foreslo derfor å redusere skalaen til 1-7. Han foreslo også at man skulle konvertere de enkeltes vurderinger av søknader til z-verdier, og separate vurderinger av de enkelte vurderingskriteriene. De to siste forslagene var identiske med de som ble foreslått for NSF.

Heller ikke her ble noen av anbefalingene tatt til følge. Noe av det mest interessante med dette er imidlertid reaksjonene han fikk fra personer som ville blitt berørt av endringene. Responsen bar preg av at de berørte personene ikke ville godta forskningsresultatene Arkes siterte, og at de oppfattet forslagene som utilbørlig inngripen i deres fagfelt. Arkes presenterer et par sitater fra de irriterte forskerne:

- "No psychologist is going to tell me how to evaluate proposals in [my field]." Denne forskeren avviser tydeligvis at det finnes noe allmenngyldig ved menneskelige vurderinger som beslutningspsykologer kan belyse. Er du ikke (for eksempel) biolog kan du ikke mene noe om vurderingskriterier for biologisk forskning.

- "Everyone can play this game and they can play it with their gut." Dette demonstrerer manglende kjennskap til forskningen i bedømming

og beslutningspsykologi. Arkes tørre kommentar er: "these fine scientists are not aware of any procedures that have improved on the gut as a decision tool".

- Arkes adferd "causes less agreement and consensus than we like to have around here." Siden vi alle er tilhengere av våre dårlige metoder vil vi ikke høre om andre og bedre metoder. Denne uttalelsen kom da Arkes argumenterte mot en administrators påstand om at holistiske vurderinger er bra ved å referere til forskningen på feltet. Fantastisk holdning hos en administrator av forskningsmidler.

- "The scientific data aren't relevant." Forskning om vuderinger generelt er ikke relevant *her*. Ingen grunn ble gitt, og Arkes har senere eksplisitt vist at, jo, forskningen generaliserer greit til domenet "evaluation of scientific proposals" (Arkes et al., 2006)

- "We don't want criteria." Disse forskerne bruker fortsatt kriterier for å vurdere prosjektforslag, enten de vil eller ei. Valget står mellom implisitte subjektive kriterier eller eksplisitte kriterier. Bruken av subjektive kriterier var akkurat det GAO hadde kritisert. Arkes kommentar er at når man forvalter store summer i forskningsmidler bør man etterstrebe å fordele disse på en mest mulig rettferdig måte, og dette oppnås bare ved bruken av eksplisitte kriterier som er kjent av søkerne.

Om man ser bort fra muligheten om at Arkes er en ufordragelig person å forholde seg til, vitner slike holdninger om alvorlige problemer med å innføre selv de enkleste tiltak for å bedre beslutninger. Andre forskningsarbeider vitner også om at det er vanskelig å innføre selv validerte beslutningsstøtteverktøy i klinisk praksis, som illustrert av van Steenkiste et als forsøk med å påvirke hjertepasienters risikoadferd (van Steenkiste et al., 2007), selv om det også finnes suksesshistorier som Waljee et als forsøk vedrørende brystkreftpasienter (Waljee et al., 2007).

Det er altså ikke nødvendigvis enkelt å innføre beslutningsstøtte i praksis. Mulige årsaker til og botemidler mot dette ses på i senere kapitler.

Del II

Diskusjon

Kapittel 4

Er beslutningsstøtte bra?

Det er viktig å skille mellom det teoretiske: "prediksjoner *kan* forbedres med statistisk prediksjon", og det praktiske "prediksjoner *bør* forbedres med statistisk prediksjon". Når jeg allikevel hevder at statistisk prediksjon bør brukes i større grad enn i dag må jeg derfor komme opp med bedre årsaker enn det gamle "fordi vi kan". Dette kan spores til skillet mellom deskriptiv og normativ forskning. Deskriptiv oppdagelse av avvik fra norm *må* ikke resultere i omfattende endringer, dette bør argumenteres for først. Jeg vil derfor gå gjennom en del grunner jeg mener taler til fordel for den økte treffsikkerheten som kan oppnås gjennom statistisk prediksjon.

4.1 Etiske betraktninger

I forhold til argumentasjon rettet mot klinikere om å ta beslutningsstøtte i bruk, har Dawes (2002, 2005) nylig trukket frem etiske problemstillinger knyttet til bruk eller ikke bruk av aktuarisk prediksjon. Hans argumentasjon går ut på at i det øyeblikket en SPR er utviklet for bruk i et spesifikt domene (og validert mot dette) så er det uetisk ikke å ta den i bruk. Når forskningen er så entydig som vi ser her, så blir bevisbyrden dyttet over på de som mener klinisk prediksjon er bedre. Da funnene er så generelle må det også presenteres evidens for hvorfor statistisk prediksjon er uegnet i hver enkelt situasjon. Dersom ingen evidens finnes på et spesifikt område tilsier forskningen allikevel at statistisk prediksjon sannsynligvis er like bra som eller bedre enn klinisk skjønn. Hvis motstandere ikke kan, eller vil, presentere forskning som slår bena vekk under aktuarisk prediksjon, og de allikevel velger å ikke benytte tilgjengelige SPR, så blir de moralsk ansvarlige for å velge suboptimale metoder. De trenger ikke basere alle prediksjoner utelukkende på SPR, men de *må* i det minste ta dem med i helhetsvurderingen.

> "Thus, a major implication of all the research is that to practice ethically, the practitioner *must* employ SPR's [...] when they are available. Moreover, the practitioner claiming to use his or her own intuition to "improve" an SPR has an ethical obligation to keep track of outcomes to see if modification really does result in improvement."
>
> — *Dawes, 2002; s. 5*

Dawes argument er som følger: Behandlere har et etisk imperativ om å predikere så godt som mulig når det de predikerer har (tildels store) konsekvenser for andre. Vi vet, gjennom forskningen, at SPR er generelt bedre, raskere og billigere (Grove et al., 2000) enn klinisk prediksjon, og at unntakene er få. Hvis man da ikke benytter seg av dem der de er tilgjengelige har man dermed med vitende og vilje benyttet en metode som sannsynligvis er dårligere, tregere og dyrere enn nødvendig, noe som er klart uetisk (Dawes, 2005).

Meehl har også argumentert for mer utstrakt bruk av beslutningsstøtte på etisk grunnlag. Som sitert i Ægisdottir et al. (2006) sier han:

> "We have no right to assume that entering the clinic has resulted in some miraculous mutations and made us singularly free from the ordinary human errors which characterized our psychological ancestors."
>
> — *Meehl, 1954; s. 28; i Ægisdottir et al. (2006)*

Et mulig motargument til disse påstandene kan være at pasienter ikke ønsker at behandlere skal bruke slike metoder. Det er en viss evidens for at dette er tilfellet (Promberger and Baron, 2006), og det kan være at den observerte motviljen mot å ta slike metoder i bruk stammer fra behandlernes innsikt i sine pasienters ønsker i så henseende. Jeg vil drøfte dette nærmere i 5.6.

4.2 Reliabilitet/standardisering

Som tidligere nevnt er det antatt at økt reliabilitet forklarer mye av årsaken til statistiske metoders økte treffsikkerhet. Dette ses på som positivt også ut fra likhetsprinsippet om lik behandling for lik tilstand uansett hvilket sykehus

en sogner til, eller hvilken lege som måtte være på jobb. Denne standardiseringsiveren har vært gjenstand for kritikk fra praktiserende klinikere som hevder at alle pasienter har rett til individualisert behandling utfra et helhetssyn, og mener at for mye kategorisering er negativt for pasientene (Berg, 1997). De som har ivret mest for standardisering har også gjerne vært administratorer og ledelse, som har sett standardisering som et verktøy for å overvåke ressursforbruk og effektivisere drift. Dette har også ført til en del skepsis og påstander om "samlebåndsproduksjon".

Det er imidlertid et faktum at mange medisinske praksiser opp gjennom tidene har vært dårlig dokumentert og hatt dårlig eller negativ effekt. Eksempler som årelating og lobotomi har ført til at medisinen nå er svært skeptisk til nye metoder før de er grundig dokumentert. Evidensbasert medisin (evidence-based medicine) har tatt over som den rådende filosofien i medisinsk praksis, og standardisering, i alle fall for forskningsformål, er essensielt for å skape statistisk dokumentasjon for nye (og eksisterende) behandlingsformer (Sim et al., 2001). Fra dette ståstedet synes det vanskelig å være motstander av statistisk prediksjon, nettopp fordi vi har så mye evidens for at det vil kunne redusere for eksempel feilbehandling.

Standardisering handler også om pasientrettigheter, og befolkningens toleranse for forskjellig praksis fra sted til sted er ikke veldig stor, som bevitnet gjennom media. I Dagsavisen 18. mai 2007 var det for eksempel et stort oppslag om at tilgangen til kreftmedisiner bestemmes av helseforetakene, og dermed varierer fra helseregion til helseregion. Med det nærmest absolutte kravet til likebehandling som hersker i Norge kan standardisering derfor ses på som en nødvendighet også uavhengig av eventuelle helsemessige gevinster som følge av økt reliabilitet.

Forutsetningen for at standardisering skal være positivt er naturligvis at det er mulig i praksis, og det er det ikke alle som mener. Marc Berg er lege og sosialantropolog, og diskuterer rasjonalisering av helsesektoren generelt (Berg, 1997). Han beskriver blant annet medisinsk praksis i lys av aktørnettverksteori. Han viser at medisinsk praksis kan ses på som et distribuert nettverk der det ikke er noen endelig menneskelig beslutningstaker, men at nettverket kontinuerlig tolker og bearbeider informasjon i den hensikt å holde pasienten innenfor et sett med eksplisitte eller implisitte normer (keeping the patients trajectory within acceptable limits). Et slikt nettverk oppfører seg ikke nødvendigvis konsekvent og informasjon er ikke noe som sekvensielt skaffes og så analyseres. Berg argumenterer for at innføring av beslut-

ningsstøttesystemer og standardisering av pasientbehandling ikke kan gjøres uten til dels store inngrep i eksisterende praksis. Innføringen av nye elementer vil føre til endringer i aktør-nettverket ved at det kommer flere aktører inn og at relasjonene mellom eksisterende aktører endres. Før innføringen av nye elementer må man derfor se på hvordan dette vil transformere eksisterende praksis. En slik beskrivelse av medisinsk beslutningstaking virker for meg som unødvendig komplisert, og bærer preg av en litt for bastant upartiskhet. Med en slik holdning kan man argumentere med at enhver endring vil kunne medføre uante konsekvenser og at det derfor er tryggest å la ting være som de er.

4.3 Konklusjon

På bakgrunn av de ovenstående argumentene, med spesiell vekt på de etiske implikasjonene rundt bruk av beslutningsstøtte, mener jeg vi nå trygt kan gå fra "kan" til "bør". Funnene i forskningen er så sterke, og implikasjonene av bruk så store, at bevisbyrden nå hviler på motstandere av klinisk beslutningsstøtte. De bør komme opp med gode valide argumenter supplert med rikelig evidens før de kan si at beslutningsstøtte ikke er bra for dem det virkelig gjelder, nemlig pasientene.

Kapittel 5

Årsaker til liten bruk av beslutingsstøtte

I den senere tid har det kommet noe forskning på årsaker til underbruk. Blant de viktigste er psykologiske faktorer som overkonfidens og etterpåklokskap, som bidrar til en oppfatning blant klinikere om at deres egne prediktive evner er bedre enn forskningen tilsier. Disse mulige årsakene diskuteres for seg i 5.7.3 og 5.7.4. Jeg vil også diskutere noen grunner foreslått av Meehl (1986) og foreslå et par andre faktorer som kan ha innvirkning, som tilgjengelighet til statistikk og en eventuell oppfattet assosiasjon mellom beslutningsstøtte og "kunstig intelligens", som feilet på 80-tallet.

5.1 Meehls grunner

Meehl foreslo 7 spesifikke årsaker til underbruk av statistisk prediksjon i sin artikkel fra 1986. Noen av disse forslagene har vært gjenstand for forskning, andre ikke:

Uvitenhet: Det er mange beslutningstakere som ikke kjenner litteraturen, det være seg dataene eller filosofien, angående statistisk prediksjon. Meehl beklager at kjennskap til grunnleggende teorier om klinisk beslutningstaking, som Bayes formel, er svært liten, selv blandt stipendiater i klinisk psykologi (i alle fall ved University of Minnesota på 80-tallet).

Trussel mot arbeidsoppgaver: Er det naturlig å tro at fagmiljøer skal omfavne hjelpemidler som gjør en del av deres arbeidsoppgaver overflødige? Naturligvis ikke. Om en klinisk psykolog bruker masse tid på

å tolke psykologiske tester liker hun ikke tanken på at statistiske prediksjonsregler er like bra (eller bedre).

Selvbilde: "Min profesjon defineres av denne typen beslutninger". Å delvis erstatte klinisk skjønn med en prediksjonsregel kan oppfattes som en trussel mot selvbildet til hele profesjonen. Som vi skal se i diskusjonen rundt overkonfidens og etterpåklokskap er vi også utstsyrt med mekanismer som beskytter selvbildet og som kan bidra til å avvise ideen om at alternativer til klinisk skjønn har noe for seg.

Teoretisk identifikasjon: "Jeg er Freudianer. Selv om jeg må innrømme at freudiansk teori ikke lar meg predikere noe av praktisk betydning om pasientene."

Humanitet: Å bruke en regresjonsligning til å predikere noe om et annet menneske reduserer dette mennesket til et datasett, eller en labrotte, og umuliggjør en holistisk vurdering av "hele mennesket". Dette "føles ikke riktig", og diskuteres nærmere i 5.5.

Misforstått etikk: I lys av det "dehumaniserende" aspektet ovenfor kan man ved sinnelagsetikk komme frem til at bruk av beslutningsstøtte er uetisk. Meehl hevder at når man fatter beslutninger om andres liv og helse er konsekvensetikk det eneste man kan rettferdiggjøre: "If I try to forecast something important [about a person] by inefficient rather that efficient means, meanwhile charging this person or the taxpayer 10 times as much money as I would need to achieve greater predictive accuracy, that is not a sound ethical practice. That it feels better, warmer or cuddlier to me as a predictor is a shabby excuse indeed." (Meehl, 1986; s. 374).

"Datafobi:" Meehl hevder han har observert en generell motstand mot ideen om at datamaskiner skal kunne være i stand til å gjøre noe bedre enn mennesker. Dette er også min erfaring når jeg prøver å fortelle folk hva jeg driver med. Når jeg refererer forskningen om at enkle lineære modeller kan være bedre prediktorer enn menneskelige vurderinger er responsen stort sett alltid noe à la "det kan ikke stemme".

Av disse faktorene synes uvitenhet å spille en stor rolle, noe som virker naturlig. Det er så vidt meg bekjent desverre ikke utført noen studier av helsearbeideres generelle kjennskap til beslutningsforskning og ei heller på hvorvidt "teoretisk tilhørighet" spiller inn. Det er imidlertid utført forskning som synes å bekrefte at Meehls "selvbilde" (Sieck and Arkes, 2005),

"humanitet" (i betydningen at beslutningsstøtte ikke "føles riktig") (Yates et al., 2003), "etikk" (Dawes, 2002; Dawes, 2005) og "datafobi" (Arkes, 2003; Arkes et al., 2007) gjenspeiler trekk ved beslutningsstøtte eller beslutningstakere som kan bidra til å forklare underbruk. Dette diskuteres nærmere i egne avsnitt.

5.2 Skillet mellom forskning og klinikk

Meehl (1986) beskriver som tidligere nevnt at det lenge har hersket et skille mellom forskning og klinisk praksis. Dette kan også gå inn under hans identifikasjon av "teoretisk tilhørighet" som en faktor med innvirkning på bruk av beslutningsstøtte. Berg (1997) har gått gjennom lederne i anerkjente medisinske tidsskrifter som JAMA (Journal of the American Medical Association) fra 50-tallet og utover, og han bekrefter et slikt skille. Forskningen ble ansett som noe som burde ha noe innflytelse på praksis, men "the art of medicine" handlet om så mye mer, og i stor grad om legens intuitive vurderinger. Berg hevder dog å ha sett en endring av denne holdningen, i retning av et ideal om en "scientist-practitioner" i tråd med fremveksten av evidensbasert medisin. Det er derfor grunn til å håpe at slik teoretisk tilhørighet til dårlig validerte teorier er på vikende front, og vil spille mindre rolle i fremtiden.

5.3 Sammenfall med AI-fiaskoen?

Statistisk prediksjon, sammen med andre typer beslutningsstøtte som ekspertsystemer, ble utviklet delvis parallelt med konseptet om kunstig intelligens (Artificial Intelligence, AI). AI-forkjemperne lovte imidlertid mer enn de kunne holde, og ut på 80-tallet mistet feltet mye kraft og troverdighet, og konsentrer seg nå om betydelig mindre ambisiøse mål enn "intelligens" (Copeland, 1993). En mulighet for den begrensede bruken av beslutningsstøtte generelt er at begrepet har blitt koblet med AI-begrepet, og at mange dermed har antatt at beslutningsstøtte generelt også mistet sin kraft på samme tid. En slik misoppfatning kan kanskje forklare noe av skepsisen mot beslutningsstøtte.

5.4 Personvern og innsamling av statistikk

En kan naturlig nok ikke bruke statistisk prediksjon dersom troverdig statistikk ikke finnes, og dette har vært en innvending mot innføring av slik beslutningsstøtte (det finnes imidlertid alternativer også her, som såkalte

bootstrapping-modeller, se Leger et al., 1992). Hvis man først må utføre et omfattende kartleggingsarbeid før man i det hele tatt kan utarbeide en SPR for testing blir bruk av statistisk prediksjon selvsagt mindre attraktivt for travle klinikere (jf. tilgjengelighet beskrevet i neste avsnitt). Statistikk finnes i mange tilfeller ikke i helsesektoren fordi data om utfall gitt diagnose og/eller behandling ikke samles på noen organisert måte, men må skaffes til veie i form av dedikerte forskningsprosjekter. I Norge har det faktisk vært forbudt, av hensyn til personvern, å samle personidentifiserbare data fra individuelle pasienter i en nasjonal database. Dette er nå til en viss grad endret med innføringen av Norsk Pasientregister (Sosial og helsedirektoratet, 2006). På sikt kan dette registeret være til stor hjelp for å skaffe objektiv baserateinformasjon om en rekke sykdomstilstander, noe som er et nødvendig skritt på veien mot gode statistiske prediksjonsregler. Registeret kan også skaffe objektiv informasjon om effekten av forskjellige behandlingsformer, og vil med tiden inneholde så store datamengder at resultatene bør generalisere godt til en rekke forskjellige populasjoner. Registeret synes å kunne bli et svært viktig verktøy for å samle informasjon om sanne prediktive variabler på en rekke felt.

5.5 Trekk ved beslutningsstøtten

Noen av årsakene til liten bruk av beslutningsstøtte kan nok også tilskrives faktorer ved selve beslutningsstøtten. Yates, Veinott og Patalano (2003) har undesøkt hva folk legger i begrepet "beslutningskvalitet", for på den måten å se om det er strukturelle faktorer ved beslutningsstøtte som kan forklare underbruk. Hvis beslutningstakere legger vekt på faktorer de mener er vesentlige for å oppnå god beslutningskvalitet, men som ikke inngår i et beslutningsstøtteverktøy, eller om de mener de faktorene beslutningsstøtteverktøyet legger vekt på er totalt irrelevante, kan dette selvsagt bidra til å forklare underbruk. Yates et al. ba forsøkspersonene tenke på beslutninger de selv hadde fattet. Forsøkspersonene i studie 1 skulle tenke på "lette" og "vanskelige" beslutninger, mens de i studie 2 skulle tenke på "gode" og "dårlige" beslutninger. Forsøkspersonene ble deretter blant annet bedt om å forklare (skriftlig) hva som gjorde at de klassifiserte beslutningene som gode, dårlige, lette eller vanskelige. Yates et al. gikk deretter gjennom svarene og lette etter fellestrekk. De identifiserte deretter en del superkategorier som svarene falt inn under:

5.5.1 Kjennetegn ved vanskelige beslutninger

Alvorlige konsekvenser: langvarige, potensielt irreversible konsekvenser, som for eksempel å måtte såre et annet menneske, bryte med egne prinsipper, eller andre alvorlige eller risikable konsekvenser.

Mange valgmuligheter: Overveldende antall valgmuligheter, eller faktorer å ta hensyn til.

Tung prosess: Ubehagelig, slitsom beslutningstaking, for eksempel under tidspress, mangel på erfaring eller usikkerhet.

Uklare utfall: Det er uklart hva konskvensene av beslutningen kan bli.

Uoversiktlige valgmuligheter: Det er vanskelig å vurdere valgmulighetene opp mot hverandre. Eksempelvis hvis forskjellige muligheter er gode på hver sine kriterier. Ingen utfall er klart best.

Uviss affekt: Beslutningstakeren vet ikke hvordan det vil oppleves å være i de situasjoner han må velge mellom.

Rådgivere: Ulike rådgivere eller anbefalinger strider mot hverandre eller mot en intuitiv vurdering.

5.5.2 Kjennetegn ved lette beslutninger

Trivielle konsekvenser: Reversible, kortvarige konsekvenser eller vinn-vinn situasjoner.

Begrensede valgmuligheter: Det er få valgmuligheter eller det er gitt hvilken man må velge. For eksempel oppmelding til obligatoriske fag.

Behagelig prosess: Minimal innsats, man "bare vet" hvilket alternativ man vil velge eller har erfaring med slike beslutninger tidligere.

Klare utfall: Det er lett å se hva konsekvensene blir, eller i alle fall å forestille seg hva de *kan* bli.

Oversiktlige valgmuligheter: Ett alternativ dominerer de andre, det vil si at det er minst like bra eller bedre enn noen av de andre på samtlige relevante faktorer.

Kjent affekt: Beslutningstakeren har opplevd konsekvensene av et slikt valg før, og vet hva de innbærer av ubehag/velvære.

Rådgivere: Det finnes klare anbefalinger, eller eventuelle rådgivere er enstemmige.

5.5.3 Kjennetegn ved "dårlige" beslutninger

Negativt utfall: Uavhengig av alt annet, så oppleves beslutningen som dårlig hvis utfallet i ettertid oppleves som dårlig. Interessant nok også hvis det dårlige utfallet var umulig å forutse.

Tap av positivt utfall: Beslutningen førte til at man tapte et utfall som senere vurderes å ha vært bra. Enten ved at det positive utfallet ville ha blitt opplevd om man hadde valgt annerledes, eller ved at man valgte bort noe som i ettertid ses å ha vært bra (som for eksempel å slå opp med en partner).

Bortfall av muligheter: Hvis beslutningen medfører begrenset valgfrihet nå eller senere, for eksempel angående yrkesvalg.

Prosess: Beslutningen anses som dårlig fordi beslutningsprosessen av en eller annen årsak vurderes som dårlig.

Affekt: Beslutningen anses som dårlig fordi man opplevde negativ affekt under eller etter beslutningstakingen.

5.5.4 Kjennetegn ved "gode" beslutninger

Positivt utfall: Positivt utfall har allerede inntruffet, eller er forventet å inntreffe. Eventuelt at den fattede beslutningen har en *tendens* til å gi positive utfall, eller i alle fall bedre utfall enn noen av beslutningsalternativene.

Unngåelse av negativt utfall: Beslutningen førte til unngåelse av et negativt utfall, eller "reddet" beslutningtakeren fra en situasjon som i utgangspunktet var negativ.

Forbedring av muligheter: Nye (presumptivt gode) muligheter er åpnet eller oppdaget som følge av beslutningen.

Prosess: Beslutningsprosessen var på en eller annen måte "god".

Affekt: Beslutningen resulterte i god affekt, eller selve beslutningen "føltes god", for eksempel ved at man gjorde "det rette".

5.5.5 Relevans for statistisk prediksjon

Yates et al. hevder at faktorer som dette må tas hensyn til ved design av beslutningsstøtteverktøy, i alle fall dersom man ønsker at beslutningstakerne spontant skal se nytten av, og ønske å bruke, beslutningsstøtte. De trekker spesielt frem at forsøkspersonene i denne undersøkelsen la stor vekt på å føle seg komfortable med beslutningsprosessen. Statistisk prediksjon kan da føre til at det er den statistiske formelen som får "æren" for beslutningen. En underliggende faktor ved en "god" prosess synes også å være at den kan rettferdiggjøres. En statistisk prediksjonsregel opererer som en "black-box", og forklaringen på hvorfor den kom fram til det svaret den gjorde blir av typen "det bare er sånn". Mangel på kausale forklaringer kan dermed være en kilde til at prosessen blir oppfattet som dårlig. Yates et al. trekker også frem det faktum at statistikk bygger på mange hendelser, mens kulturen i vestlige land er svært individualistisk:

> "That approach ignores the uniqueness that is prized almost as a moral imperative in individualist cultures like that of the United States."
>
> — *Yates et al., s. 38*

Manglende evne til å ta hensyn til unike "broken leg-argumenter" trekkes også frem som en faktor som påvirker beslutningstakeres vurdering av kvaliteten på en beslutning fattet ved hjelp av statistisk prediksjon. Statistisk prediksjon benytter kun de variablene som har demonstrert validitet og ignorerer kanskje mye informasjon beslutningstakeren synes *burde* tas med i betraktningen, fordi de ikke er klar over at dette ikke øker treffsikkerheten.

Det kan altså virke som om enkelte trekk ved statistisk prediksjon strider mot en del av det mennesker oppfatter som sentrale trekk ved en "god" beslutning. Implikasjonene dette har for mulige forbedringer av beslutningsstøtte diskuteres nærmere i 6.1

Konkrete faktorer som brukervennlighet og tilgjengelighet spiller selvsagt også inn. Meg bekjent er det ingen som har undersøkt beslutningsstøttesystemer med tanke på slike faktorer. Sannsynligheten for at brukervennlighet og tilgjengelighet i de fleste tilfeller er tatt hensyn til burde dog være relativt høy med tanke på at både beslutningsforskning og forskning på "human factors" begge har sitt utspring i psykologien. Det er i alle fall å håpe at forskere som selv beklager at det er for lite kunnskap om deres felt (Meehl, 1986; Arkes, 2003) ikke selv har oversett relevant psykologisk forskning.

5.6 Trekk ved pasientene

De store forkjemperne for økt bruk av statistisk prediksjon, som Dawes og Arkes, har argumentert for beslutningsstøtte med grunnlag i at det later til å kunne øke den prediktive treffsikkerheten hos kliniske profesjonsutøvere. De har imidlertid ikke nevnt med et ord hvordan de mener *pasientene* vil reagere på beslutningsstøtte. Dette må jo sies å være en sentral del av diskusjonen om beslutningsstøtte i helsesektoren, og det finnes heldigvis en del forskning også på dette emnet.

5.6.1 Menneske-maskin-interaksjon - stoler pasienter på datamaskiner?

Når vi snakker om statistisk prediksjon og beslutningsstøtte er dette ofte i forbindelse med et datamaskinbasert system (for eksempel Grundmeier & Johnson, 1999). En diskusjon om pasienters forhold til beslutningsstøtte bør derfor inneholde en diskusjon av pasienters forhold til datamaskiner. Før jeg kommer inn på funn på dette konkrete emnet er det imidertid nødvendig med en litt bredere innføring i menneskers holdninger til feil.

Feil kan stort sett alltid oppstå i alle typer systemer, også i helsevesenet. Mulighetene for å redusere feil er en sterk motivator for økt bruk av beslutningsstøtte. Mennesker reagerer derimot ikke likt på alle typer feil, og det er grunn til å tro at pasienter vil reagere ulikt på feil som oppstår på grunn av feilvurdering fra en lege og på feil som oppstår på grunn av teknologiske systemer som datamaskinbaserte beslutningsstøttesystemer (Naquin and Kurtzberg, 2004). En av mekanismene bak slike ulike reaksjoner er kjent som "kontrafaktisk tenkning" i beslutningslitteraturen (Roese, 1997).

Kontrafaktisk tenkning er en teori som forklarer menneskelige avvik fra rasjonalitet utfra hvor lett det er å se for seg alternative utfall. Hvis vi har tatt et valg som ledet til et negativt utfall avgjøres nivået av anger (regret) av hvor lett det er å konstruere alternative (kontrafaktiske) historier med et annet og bedre utfall. Hvis det er lett for meg å forestille meg hvordan det negative utfallet kunne vært forhindret vil jeg klandre meg selv mer enn hvis det negative utfallet fremstår som uunngåelig.

Kontrafaktisk tenkning ("Hvis jeg bare hadde gjort slik i stedet") antas derfor å ha en sentral rolle i hvordan vi reagerer på feil som attribueres til henholdsvis menneskelige og teknologiske årsaker. Naquin og Kurtzberg (2004)

fant ut at dersom en ulykke ble attribuert til teknologiske årsaker (togulykke som følge av signalfeil) anså forsøkspersonene togselskapet som mindre ansvarlig enn hvis den samme ulykken ble attribuert til menneskelig svikt (togføreren kjørte på rødt). Naquin og Kurtzberg fant også en sammenheng mellom feilattribusjonen og antallet kontrafaktiske tanker forsøkspersonene oppga da de ble bedt om å forestille seg hvordan ulykken kunne vært unngått. Den menneskelige togføreren anses å være i stand til å handle anderledes enn han gjorde, og har derfor ansvaret for å handle riktig. Lyssignalet derimot, kan ikke holdes ansvarlig for sine handlinger.

Maskiner kan dermed ikke ta ansvar, mens mennesker kan. Dette spiller igjen inn på pasienters beslutningstaking. Promberger og Baron (2006) undersøkte pasienters reaksjon på et datamaskinbasert beslutningsstøtteverktøy som presenterte pasientene for konkrete anbefalinger, versus den samme anbefalingen fra en lege. Hypotesen var basert på at pasienter som må velge behandlingsform er motivert av mer enn bare det medisinske utfallet, de er også opptatt av å minimere sin egen skyldfølelse dersom noe skulle gå galt. Om de får en anbefaling fra en lege, og følger denne, er det legens skyld om ting går galt. Om man derimot følger rådet fra en maskin er det ens eget ansvar. Promberger og Barons hypotese gikk ut på at pasienter dermed ville være mer tilbøyelige til å følge anbefalinger fra leger enn fra datamaskiner ut fra et ønske om å flytte ansvaret for eventuelle negative utfall vekk fra seg selv.

Promberger og Baron testet denne hypostesen i et scenario med fire mulige utfall. To grupper ble presentert med en oppsummering av symptomer samt en anbefaling om de burde gjennomgå en operasjon, og denne anbefalingen kom fra enten en lege eller et dataprogram. For de to andre gruppene var anbefalingen at de ikke burde gjennomgå noen operasjon, og denne kom også fra enten en lege eller et dataprogram. De fant en effekt for begge anbefalingene. Pasienter var mer tilbøyelige til å gjennomgå operasjon når anbefalingen kom fra en lege, og de var også mer tilbøyelige til å avstå fra operasjon når dette ble anbefalt av en lege. De testet også om dette hadde sammenheng med attribusjon av ansvar. De fant at pasientene alltid følte mer ansvar for en beslutning hvis de brøt med en konkret anbefaling, og denne ansvarsfølelsen ble større ved å bryte en anbefaling fra en lege enn ved å bryte en anbefaling fra et datasystem.

I en separat studie (i samme artikkel) testet de så om det var forskjell i beslutningsinnstilling (decision attitude) og tillit (trust) avhengig av om

forsøkspersonene fikk presentert en anbefaling og en avgjørelse fra leger eller datasystemer. Det var altså fire mulige utfall: En beslutning presentert av en lege eller et datasystem, og en anbefaling presentert av en lege eller et datasystem. Hypotesene var som følger:

Hypotese 1: Forsøkspersonene vil være mer beslutningssøkende (decision seeking) når de samhandler med et datasystem enn når de samhandler med en lege. Det vil si at de i større grad foretrekker å treffe egne valg når alternativet er at et datasystem tar avgjørelsen enn dersom en lege er den alternative beslutningstakeren.

Hypotese 2: Forsøkspersonene vil være mer beslutningssøkende når de får beskjed om å godta en ferdig avgjørelse enn når de blir rådet til å godta en anbefaling (avgjørelsene ble presentert som "du kan nyte godt av behandling A eller B. behandling A er ikke tilgjengelig der du bor. Legen avgjør at du får behandling B.")

Hypotese 3: Forsøkspersonene vil oppgi høyere tillit til en beslutning eller anbefaling når den kommer fra en lege enn når den kommer fra et datasystem.

De fant effekter for alle de fire utfallene hva gjelder beslutningsinnstilling, men bare en effekt var statistisk signifikant. Det var kombinasjonen av hypotese 1 og 2, forsøkspersonene var beslutningsaversive (det vil si innstilt på å følge anbefaling/avgjørelse) når de fikk en anbefaling fra en lege. Anbefaling fra datasystem og avgjørelse fra enten datasystem eller lege resulterte ikke i noen signifikant trend. For tillit fant de at forsøkspersonene hadde signifikant høyere tillit til beslutningskvaliteten til både anbefalinger og avgjørelser tatt av leger enn av datasystemer. De fant også at kvaliteten på en anbefaling generelt ble sett på som noe høyere enn kvaliteten på en avgjørelse, men denne effekten var liten.

Denne forskningen viser altså at (hypotetiske) pasienter er mer innstilt på å følge anbefalinger fra leger enn fra datasystemer. Dette henger sammen med attribusjon av ansvarlighet for beslutningen, at pasientene i større grad ønsker å fatte beslutningene selv når den alternative beslutningstakeren er et datasystem enn om det er en lege og at de har høyere tillit til beslutningskvaliten hos en lege enn et datasystem. Denne forskningen er ny, og i hvor stor grad det er mulig å påvirke pasienters innstilling til datasystemer er foreløpig lite undersøkt. Promberger og Baron ønsker å undersøke dette nærmere i sin videre forskning, og det er dermed håp om at vi kan få svar på disse spørsmålene i en ikke alt for fjern fremtid.

Promberger & Barons (2003) funn er svært relevante både for å forklare underbruk av beslutningsstøtte rettet mot klinikere (det kan være at klinikerne tar hensyn til pasientenes holdninger til beslutningsstøtte) og med tanke på det økende antallet beslutningsstøttesystemer som retter seg mot pasienter i stedet for helsepersonell.

5.6.2 Mellommenneskelige forhold - stoler pasienter på leger?

Pasienters holdning til beslutningsstøtte for egen del er relevant, men det er også aktuelt å undersøke pasienters holdninger til legers bruk av beslutningsstøtte. Hvis pasienter ikke har tillit til beslutningskvalieten til beslutningsstøttesystemer, hva mener de da om leger som benytter slike systemer? Dette er undersøkt av Pezzo & Pezzo (2006) og Arkes, Shaffer og Meadow (2007). Pezzo og Pezzo undersøkte hvordan bruken av et beslutningsstøttesystem påvirket folks oppfatning av ansvarlighet etter en hypotetisk medisinsk feilvurdering. Arkes, Shaffer og Meadow undersøkte folks oppfatning av en leges dyktighet basert på bruk av beslutningsstøtte uavhengig av utfall.

I Pezzo og Pezzos studie 1 undersøkte de effekten av å bruke et beslutningsstøttesystem på forsøkspersonenes oppfatning av legens kompetanse, og på graden av uaktsomhet hvis utfallet ble negativt. Det var fire grupper i eksperimentet. To grupper der legen avdekket et potenisielt alvorlig problem, enten med eller uten beslutningsstøtte, og to grupper der problemet ikke ble oppdaget, også her med eller uten beslutningsstøtte.

Kompetanse: For de to gruppene med positivt utfall ble legen ansett som mer kompetent uten bruk av beslutningsstøtte. For de to gruppene med negativt utfall ble legen derimot ansett som mer kompetent ved bruk av beslutningsstøtte. Bruk av beslutningsstøtte beskyttet altså til en viss grad ved negativt utfall, men på bekostning av legens oppfattede kompetanse hvis utfallet var positivt.

Uaktsomhet: For de to gruppene med negativt utfall fant de at bruk av beslutningsstøtte førte til signifikant reduksjon av forsøkspersonenes vurdering av uaktsomhet.

I studie 1 fulgte legen alltid anbefalingen fra beslutningsstøttesystemet. Pezzo og Pezzo ønsket også å undersøke effekten av å bruke et beslutningsstøttesystem

der man bryter med systemets anbefalinger. I studie 2 undersøker de legestudenters og legfolks (disse var som vanlig også studenter...) vurdering av ansvar for feil, kompetanse og den generelle hensiktsmessigheten ved bruk av beslutningsstøtte (appropriateness of using a computerized decision aid in general) ved fire scenarier, alle med negativt utfall. Legen var enten enig med og fulgte anbefalingen fra beslutningsstøtten, var uenig, men fulgte anbefalingen, var uenig og brøt med anbefalingen, eller benyttet ikke beslutningsstøtte i det hele tatt. Designet ble dermed 2 (medisinstudent vs. ikke-medisinstudent) x 4 (enig, følge, bryte, kontroll).

Ansvar for feil: De fant at legfolk generelt holdt legen mer ansvarlig for feil enn medisinstudenter. De fant også at legen ble ansett som mer ansvarlig for feil hvis han brøt med anbefalingen fra beslutningsstøttesystemet. Om han fulgte anbefalingen eller ikke brukte beslutningsstøtte i det hele tatt ga omtrent like utslag.

Kompetanse: Legfolk vurderte legen som signifikant mindre kompetent enn legestudentene. Begge gruppene vurderte legen som mindre kompetent hvis han brøt med beslutningsstøtten eller ikke brukte beslutningsstøtte i det hele tatt. Det spilte ingen rolle om legen var enig eller uenig i beslutningsstøttens anbefaling så lenge han fulgte den.

Hensiktsmessighet: Begge grupper mente det var moderat hensiktsmessig å bruke datamaskinbasert beslutningsstøtte generelt, men medisinstudentene mente det var mer hensiktsmessig enn legfolkene.

Medisinstudentene hadde imidlertid også blitt bedt om å anslå hvordan de trodde legfolk ville svart i de samme situasjonene. Svarene på dette stemte godt overens med de faktiske svarene fra legfolkene, noe som indikerer at den effekten Pezzo og Pezzo fant ikke er ukjent for medisinstudenter, og dermed også sannsynligvis leger. Igjen ser vi altså at legers underbruk av beslutningsstøtte kan henge sammen med deres kjennskap til at pasienter ikke har like stor tillit til dette som dem selv. Pezzo & Pezzo viste imidlertid at legene som brøt med beslutningsstøtteverkøyet ikke ble oppfattet mer negativt enn de som ikke benyttet noen form for beslutningsstøtte. Bruk av beslutningsstøtte kan således til en viss grad beskytte mot oppfatning av uaktsomhet etter feildiagnostiseringer, selv der legen følger beslutningsstøtteverktøyets råd mot sin egen (korrekte) vurdering.

Pezzo og Pezzo undersøkte kun vuderinger av kompetanse etter negative utfall. Det er også ønskelig å se på vurderingen av kompetanse etter positive

utfall, og det er det Arkes, Shaffer og Meadow har gjort. De utførte fire eksperimenter, to med "vanlige studenter", én med medisinstudenter og én med reelle pasienter fra University of Ohios helsetjeneste. Eksperimentene var like for de vanlige studentene og de reelle pasientene, mens medisinstudentene fikk en noe annen variant. For alle fire gruppene var det en signifikant forskjell i vurderingen av legens diagnostiske kompetanse avhengig av om de brukte beslutningsstøtte eller ikke. Bruk av beslutningsstøtte førte i alle tilfeller til at legen ble vurdert som mindre kompetent enn leger som ikke benyttet beslutningsstøtte. Det eneste som til en viss grad motvirket dette var når beslutningsstøtteprogrammet var utviklet ved et kjent prestisjesykehus (Mayo clinic).

Vi ser altså at bruk av beslutningsstøtte fører til mindre varians i vurderingen av legens kompetanse. De blir vurdert mindre negativt ved negative utfall, men også mindre positivt ved positive utfall. Den foreslåtte årsaken til dette er at en del av "æren" for de negative eller positive utfallene tilskrives beslutningsstøtteverktøyet. Det kan også være at leger som bruker beslutningsstøtte fremstår som mer usikre enn de som ikke bruker det, og at de dermed anses som mindre kompetente. Jeg kommer tilbake til implikasjonene ved pasienters holdninger til beslutningsstøtte i neste seksjon.

5.7 Trekk ved helsearbeiderne

5.7.1 Menneske-maskin-interaksjon - Stoler *leger* på datamaskiner?

Som tidligere nevnt er det ikke alltid så lett for mennesker å velge det vi objektivt vet er lurest. Bruk av rusmidler foreksempel, er et stjerneeksempel på at rasjonalitet alene ikke er tilstrekkelig til å forklare menneskelig adferd. Hvis kliniske beslutningstakere opplever negativ affekt ved tanken på å bruke beslutningsstøtte kan dette være en medvirkende årsak til funn som dem beskrevet av Corey & Merenstein (1987, i Arkes, 2007). De fant som tidligere nevnt at et system som ville redusert falske positive diagnoser av akutt hjertesykdom fra 71% til 0% kun ble benyttet i 2,8% av aktuelle tilfeller. Betyr dette at legene ved dette sykehuset, og kanskje leger generelt, hater tanken på å bruke beslutningsstøtteverktøy?

Det siste er i alle fall ikke tilfelle i følge Grundmeier & Johnson (1999). De utførte en spørreundersøkelse for å avdekke holdningene leger ved John

Hopkins Hospital og George Washington University Medical Center hadde til kliniske beslutningsstøttesystemer (Clinical Decision Support Systems, CDSS). Svarprosenten var bare 27%, men de fikk fortsatt inn 209 svar, som burde være nok til å gi en viss pekepinn om holdninger. Legene som svarte var allmennleger, psykiatere, pediatere, gynekologer, generelle kirurger og anestesileger. De fant at 63% var enige eller veldig enige i at CDSS vil forbedre kvalitet (quality of care), og 52% var enige eller veldig enige i at CDSS vil redusere antallet feilmedisineringer (adverse drug effects). På spørsmål om CDSS vil påvirke produktivitet eller autonomi var de nøytrale. Legene var også for et CDSS som advarte dem om mulige negative utviklinger hos pasienten (for eksempel hypokalemi) så lenge advarslene hadde en sann prediktiv verdi (PPV) på mer enn 67% (det vil si at de ikke ville tolerere mer enn 33% falske positive). Grundmeier & Johnson konkluderer med at et godt designet CDSS med regler som overstiger en eksplisitt definert PPV kan aksepteres av klinikere.

Vi ser dermed at dersom populasjonene til Grundmeier & Johnson (1999) og Corey & Merenstein (1987) ikke er fullstendig ulike så er ikke klinikernes holdninger til CDSS alene nok til å forklare underbruken av CDSS der de er tilgjengelige. Et lite aber ved denne konklusjonen er at Grundmeier og Johnson fant en sammenheng mellom holdninger til Physician Order Entry (ut fra kontekst antatt å være tilsvarende Elektronisk Pasientjournal, EPJ) og holdninger til CDSS. Dette kan igjen være et uttrykk for holdninger til datamaskiner i klinisk praksis generelt, og dette kan godt ha endret seg fra 1987 til 1999. Grundmeier og Johnson kommenterte også at deres resultater kan være påvirket av at en uforholdsmessig høy andel av legene som svarte er vant med datamaskiner (computer savvy).

5.7.2 Pasienters holdninger til beslutningsstøtte som mediator for legenes underbruk.

Som vi har sett er det evidens for at pasienter er skeptiske til beslutningsstøtte generelt, og at de også vurderer leger som bruker beslutningsstøtte som mindre kompetente. Begge disse faktorene kan være medvirkende årsaker til dokumentert underbruk av beslutningsstøtte. Om leger er klar over sine pasienters holdninger til beslutningsstøtte kan en årsak til underbruk være at legene tar hensyn til pasientenes følelser, og at legene ønsker at pasientene skal være mest mulig fortrolige med den metoden som blir brukt for å komme frem til deres diagose og behandling. Pezzo & Pezzo (2006) fant jo at medisinstudenter var klar over at pasienter ville være mer skeptiske til

beslutningsstøtte enn det de selv var, så det er en viss evidens for dette.

Det er heller ikke unaturlig å anta at legene er klar over at de vil bli vurdert som mindre kompetente hvis de bruker beslutningsstøtteverktøy under en konsultasjon. Det er dermed forståelig om leger kvier seg for å bruke slik beslutningsstøtte, da de naturligvis ønsker å fremstå best mulig for pasientene. Både av hensyn til legens rykte, og av hensyn til tillit i lege-pasientforholdet. Hvis bruk av beslutningsstøtte får legen til å fremstå som usikker vil dette naturligvis kunne påvirke denne tilliten.

5.7.3 Overkonfidens

Det er altså demonstrert overveldende evidens for at beslutningsstøtte kan ha en plass i klinisk praksis. Klinikere er heller ikke spesiselt negative til innføring av beslutningsstøttesystemer såfremt de har demonstrert validitet og ikke overvelder klinikeren med advarsler om usannsynlige hendelser (Grundmeier and Johnson, 1999). Pasienter har en viss skepsis (Promberger and Baron, 2006; Arkes et al., 2007), men dette alene er neppe nok til å å forklare underbruk av beslutningsstøtte der det er tilgjengelig. Funn i beslutningspsykologi kan ikke bare brukes til å forklare hvorfor beslutningsstøtte kan øke prediktiv treffsikkerhet, men også til å forklare underbruk. En effekt som er foreslått å ha en sentral betydning for underbruk av beslutningsstøtte er overkonfidens. Dette går ut på at (mentalt friske) mennesker har en tendens til å anse sine egne evner som over gjennomsnittet (for en innføring i overkonfidens, se Plous, 1993). Når de da får presentert forskning som "kliniske beslutningstakere har på dette feltet en gjennomsnittlig treffsikkerhet på 73%, mot 85% hvis de konsekvent følger rådene fra dette beslutningsstøtteverktøyet" vil mange, sannynligvis de fleste, tenke at *deres* treffsikkerhet ligger over dette gjennomsnittet og at de derfor ikke vil ha nytte av beslutningsstøtten.

Folks tendens til å overvurdere sine egne prediktive ferdigheter versus et beslutningsstøtteverktøy er undersøkt av Sieck og Arkes (2005). De undersøkte gjennom tre eksperimenter spesifikt om overkonfidens var relatert til bruk av beslutningsstøtte gjennom å se om reduksjon av overkonfidens førte til økt bruk av beslutningsstøtten. Disse eksperimentene avdekket såpass interessante funn at det er verdt å gå gjennom dem i detalj. Forsøkspersonene i alle tre eksperimentene skulle vurdere om potensielle jurymedlemmer var tilhengere av aktiv dødshjelp basert på demografiske data (Alder, politisk tilhørighet, alkoholforbruk, religiøsitet og holdning til førekteskapelig sex). Baseraten for

tilhengere av aktiv dødshjelp var 61.7%. I alle tre eksperimentene var det en treningsfase og en testfase. Halvparten av forsøkspersonene fikk varierende former for feedback under eller etter treningsfasen, kontrollgruppen fikk ingen feedback. Beslutningsstøtte var tilgjengelig for noen av forsøkspersonene, da i både trenings- og testfasene, og det var valgfritt om de ville benytte denne. Eksperimentet foregikk via et dataprogram, og beslutningsstøtten inngikk i dette programmet. Forsøkspersonene måtte trykke på en knapp på skjermen for å få tilgang til prediksjonsregelens utfall i hver enkelt sak. De måtte så vurdere om de ville følge regelen eller bryte med den. De ble også bedt om å angi sin holdning til statistikk generelt (attitudes towards statistics, ATS).

Eksperiment 1

Eksperiment 1 skulle teste hvorvidt enkel feedback ville redusere overkonfidens, og eventuelt påvirke bruk av beslutningsstøtte. Forsøkspersonene ble først kjørt gjennom en rekke på 120 demografiske datasett om personer (treningsfase), fikk en pause og ble så kjørt gjennom en ny rekke på 60 saker (testfase). Halvparten fikk feedback på hvorvidt de hadde rett etter hver sak, de andre fikk ingen feedback. To grupper hadde også tilgang til en statistisk formel som de ble fortalt hadde en treffsikkerhet på 77%, en av disse gruppene fikk feedback, den andre ikke. Etter testen ble deltakerne bedt om å oppgi hvor ofte de mente de hadde oppgitt samme svar som formelen. De ble så bedt om å anslå i hvor mange prosent av disse sakene de trodde de hadde oppgitt rett svar. De ble også bedt om å anslå hvor ofte de hadde rett i de tilfellene der de var uenige med formelen. Resultatene viste at feedback ikke hadde noen effekt for treffsikkerhet, henholdsvis 61% for "no-feedback" og 62% for feedback. Tilgang til den statistiske formelen økte treffsikkerheen marginalt, til 65% både for no-feedback og feedback. Den statistiske formelen ble lite brukt, i "feedback-gruppen" ble resultatene fra formelen undersøkt i 30% av sakene, i "no-feedback-gruppen" i 43% av sakene (denne forskjellen anses av Sieck og Arkes som ikke signifikant, uten at de har oppgitt noen årsak til dette. Min antakelse er at standardavviket var svært stort). Det var en relativt høy korrelasjon mellom holdning til statistikk og bruk av beslutningsstøtten. Det mest slående funnet var imidlertid forsøkspersonenes konfidens i sine egne vuderinger der de visste at de var uenige med formelen. Selv om de visste at formelen hadde rett i 77% av tilfellene oppga både feedback-gruppen og no-feedback gruppen at de trodde de hadde rett i rundt 50% av de sakene der de var uenige med formelen. Samtlige grupper ville gjort det bedre om de hadde basert seg utelukkende på formelen. Enkel feedback reduserte heller ikke overkonfidensen. Sieck og Arkes foreslår på bakgrunn av dette eksperimentet to faktorer som kan ha innvirkning på bruk av be-

slutningsstøtte. (1) Generell holdning til statistikk. (2) Forhøyet tro på egne vurderinger (overkonfidens).

Eksperiment 2

I eksperiment 2 forsøkte Sieck og Arkes å redusere overkonfidens ved å fokusere på to forskjellige typer feedback som tidligere er vist å ha effekt. Den ene metoden var "performance monitoring (PM)" der forsøkspersonene eksplisitt ble bedt om å kontrollere sine egne ferdigheter opp mot formelen, med spesiell vekt på tilfeller der de i utgangspunktet var uenige med den (de ble bedt om å sjekke sine svar mot formelens svar 3 ganger i løpet av treningsfasen, og svare ja eller nei på hvorvidt de hadde gjort det bedre enn formelen). Den andre var "calibration feedback (CF)" der forsøkspersonene fikk presentert grafer over faktisk treffsikkerhet og sin egen vurdering av treffsikkerhet etter treningsfasen. Det ble forklart at hvis deres egen vurdering av treffsikkerhet lå over grafen for faktisk treffsikkerhet indikerte dette overkonfidens. Grafene ble presentert for alle sakene som helhet, for de sakene der de hadde oppgitt samme svar som formelen og for de sakene der de var uenige med formelen. Begge gruppene hadde tilgang til formelen i begge faser, og begge gruppene fikk i tillegg samme type individuell feedback for hver sak som de fikk i eksperiment 1. Resultatene var imidlertid nedslående, da treffsikkerhet og overkonfidens var stort sett identisk for alle gruppene på 67-68% (dog noe bedre enn det som ble oppnådd i eksperiment 1). De klarte heller ikke her å redusere overkonfidens i testfasen, selv med grundig og variert feedback etter treningsfasen. Den eneste effekten de fant var at de i PM-gruppen aksesserte formelen oftere, og var svært marginalt flinkere til å bruke den (men det var ingen signifikant forskjell i faktisk treffsikkerhet). Holdning til statistikk hadde betydning for de i CF-gruppen, men denne effekten forsvant i PM-gruppen. Effekten var imidlertid signifikant for gruppen som helhet ($r = 0.21$).

Eksperiment 3

I eksperiment 3 forsøkte Sieck og Arkes nok en gang å redusere overkonfidens, denne gangen ved å gi calibration feedback som i eksperiment 2, men forsøkspersonene skulle her også memorisere informasjonen fra grafene og svare på spørsmål for å bekrefte at de hadde forstått hva dette betød med tanke på overkonfidens. De kunne gå tilbake og sjekke grafene, men grafene og spørsmålene var på forskjellige skjermbilder. Håpet var at denne metoden, kalt enhanced calibration feedback (ECF) skulle sørge for en grundigere bearbeidelse av informasjonen i grafene, og på den måten redusere overkonfidens.

De ønsket imidlertid ingen økt treffsikkerhet fra dette alene, og kontrollerte derfor for dette. ECF ga en liten reduksjon i overkonfidens, men påvirket ikke treffsikkerhet i det hele tatt når ingen formel var tilgjengelig. I dette eksperimentet fikk forsøkspersonene ikke tilgang til den statistiske formelen før i testfasen, og de fikk ikke individuell feedback på hver sak undeveis i treningsfasen. Treningsfasen ble også redusert fra 120 til 60 saker av praktiske årsaker. Resultatene viste en liten, men signifikant økning i treffsikkerhet, og en tilsvarende reduksjon av overkonfidens for ECF-gruppen. Viktigst for Sieck og Arkes hovedhypotese (overkonfidens bidrar til underbruk av beslutningsstøtte) var at ECF-gruppen undesøkte formelen oftere (68% av gangene, mot 48% i kontollgruppen), og når de undersøkte formelen var de mer tilbøyelige til å følge den (87% av gangene, mot 78% i kontrollgruppen). Selv om de fant at ECF førte til redusert overkonfidens var denne fortsatt signifikant, og faktisk gjennomsnittlig treffsikkerhet for ECF-gruppa var bare 68%. De ville altså også her ha gjort det bedre ved å følge formelen slavisk. De brukte også formelen mye sjeldnere enn de kunne gjort. Et interessant funn i eksperiment 3 er også at effekten av "holdning til statistikk" forsvant (korrelasjonen var fortsatt positiv, men ikke signifikant).

Sieck og Arkes fant altså at overkonfidens har en sammenheng med bruk av beslutningsstøtte. De fant imidlertid også at overkonfidens er svært robust, og folk har en tendens til å stole på mageføelse også når denne strider mot en validert beslutningsregel. Yates (2003) har funnet at mennesker generelt er fornøyd med de besltuningene de fatter, og i ettertid mener de var gode. Når de da har muligheten til å benytte et beslutningsstøttesystem ser de ikke nødvendigheten, på grunn av overkonfidens angående sine egne evner. Utfordringen er altså todelt. For det første må man prøve å overbevise klinikere til å i det hele tatt vurdere beslutningsstøtten, og så må man få dem til å ta hensyn til beslutningsstøttens anbefalinger.

5.7.4 Etterpåklokskap

Etterpåklokskap vil si at våre sannsynlighetsvurderinger om hendelser i fortiden i stor grad påvirkes av hva som faktisk skjedde (Plous, 1993). Hendelser fremstår dermed i ettertid som langt mindre usikre enn de i virkeligheten var. Yates, Veinott og Patalano (2003) bruker etterpåklokskap til å forklare hvorfor forsøkspersonene i deres eksperimenter rapporterte en svært høy grad av tilfredshet med beslutninger de tidligere hadde fattet, også beslutninger de selv hadde beskrevet som vanskelige. En slik høy grad av tilfredshet vil sannsynligvis bidra til å underbygge bildet av en selv som en habil beslutningstaker.

Kapittel 6

Tiltak for å øke bruk av beslutningsstøtte

Det er altså mange mulige årsaker til liten bruk av statistisk prediksjon. Kan så noen av disse årsakene motvirkes? Det er flere innfallsvinkler til dette. En kan se på mulighetene til å forbedre selve beslutningsstøtten, påvirke beslutningstakerne, eller påvirke organisatoriske faktorer.

6.1 Forbedring av beslutningsstøtten

Vi skal først se på faktorer ved selve beslutningsstøtten. Manipulasjon av beslutningsstøtten som forsøk på å øke bruk har vært forholdsvis lite undersøkt, men Kaplan, Reneau & Whitecotton (2001) har så vidt sett på et par faktorer. Kaplan et al. undersøkte virkningen av å manipulere informasjonen om beslutningsstøttens prediktive valididet, og effekten av å involvere beslutningstakerne i utformingen av beslutningsstøtten. Noen av funnene fra Yates et al (2003) er også relevante her.

6.1.1 Informasjon om Positiv Prediktiv Verdi (PPV)

Kaplan et al. avdekket at bruk av beslutningsstøtten økte markant hvis man ikke oppga beslutningsstøttens PPV eksplisitt. Forsøkspersonene var erfarne revisorer som skulle predikere 16 selskapers antatte kredittverdighet (4 mulige klassifiseringer). De hadde tilgjengelig 3 prediktive variabler og en statistisk prediksjonsregel. Gruppen ble delt i to, og den ene gruppen fikk oppgitt at SPRen hadde en treffsikkerhet på 50% (det vil si dobbelt så bra som å velge på måfå). Den andre gruppen fikk bare presentert SPRen, uten noen informasjon om treffsikkerhet. Det ble også kontrollert for personlighets-

trekket "locus of control", som beskriver i hvor stor grad en person mener han kan kontrollere omgivelsene rundt seg. Har man intern locus of control ser man på seg selv som i stor grad å ha kontroll over omgivelsene. Ekstern locus of control vil si at man i stor grad mener at omgivelsene kontrollerer en selv. Kaplan et al fant at den gruppen som hadde fått informasjon om treffsikkerhet fattet samme beslutning som SPRen 61.1% av gangene, mens de som ikke hadde fått informasjon om PPV var enige med SPRen hele 77.9% av gangene. Som forventet var bruk høyere for de med ekstern locus of control enn for de med intern locus of control, noe som indikerer at individuelle forskjeller i personlighetstrekk også kan spille inn.

Kaplan et al. forklarer resultatet med en form for overkonfidens. De hevder at overkonfidens ikke bare gjelder mennesker, men at vi også gjerne er overkonfidente på vegne av beslutningsstøtten. Å gi eksplisitt informasjon om beslutningsstøttens PPV reduserer dermed overkonfidensen for beslutningsstøtten, men ikke for oss selv. Forskjellen på intern og ekstern locus of control kan skyldes forskjeller i overkonfidens attribuert til en selv eller omgivelsenes evne til å predikere. Denne automatiske kalibreringen av statistisk prediksjon nevnes også av Dawes (2005), om enn i et noe annet lys:

> "Finally, SPRs have a particular virtue that human intuition does not; by specifying how well a prediction can be made, they automatically specify how *badly* it is made. [...] Ironically, the automatic specification of how poorly a prediction is made yields a temptation to believe that because it *could*, hypothetically anyway, be made better, the relevant SPR should be abandoned in preference to some unproven method—most often our own intuition, which we have found to be worse. As on other occasions, ethical behavior involves resisting temptation."
>
> — *Dawes, 2005; s. 1254*

Kan vi så bruke slik selektiv tilbakeholdelse av prediktiv informasjon for potensielt å øke bruk av statistisk prediksjon i helsesektoren? Det forekommer meg tvilsomt at klinikere vil ta i bruk beslutningsstøtte uten oppgitt validitet. Som vi husker fra Grundmeier & Johnson (1999) forlangte legene en PPV på mer enn 67% før de i det hele tatt ville akseptere at et beslutningsstøttesystem grep inn i arbeidet deres. Det man imidlertid kan gjøre er å legge liten vekt på PPV når man introduserer beslutningsstøtten, minst mulig "salience", for på den måten å nyttiggjøre seg tendensen vi har til å dele vår overkonfidens med beslutningsstøtten.

6.1.2 Medvirkning

Kaplan et al. foreslo også brukermedvirkning som en faktor som kan øke bruk. For å teste dette laget de en modifisert versjon av "PPV-eksperimentet". Eksperimentgruppen i dette eksperimentet kunne selv velge hvilke 3 av 5 prediktive variabler de ville at den statistiske prediksjonsregelen skulle benytte (regelen som faktisk ble benyttet var den samme som i forrige eksperiment, medvirkningen var således komplett illusorisk) for på den måten å føle at de "var med" på beslutningen. Også her kontrollerte de for locus of control, og hypotesen var at medvirkning ville øke bruk av SPRen mer for de med intern locus of control enn de med ekstern. Effekten de fant var dramatisk. For de med ekstern locus of control førte økt medvirkning til en økning i enighet med SPRen fra 50.9% til 60.9%. Hos de med *intern* locus of control førte økt medvirkning til en økning i enighet fra 40% til 69.8%.

Virkningen av en slik illusorisk følelse av medvirkning beskrives også av Berg (1997), som beskriver et ekspertsystem for diagnostisering av appendisitt (blindtarmbetennelse). Da legene som skulle bruke systemet så det første gang insisterte de på at variablene "puls" og "temperatur" skulle inkluderes i systemet. Utviklerne av systemet hadde undersøkt den prediktive validiteten av disse variablene og funnet ut den var lik null. De la allikevel inn felter for temperatur og puls i dataprogrammet, men informasjonen i disse feltene ble ikke brukt for å komme frem til diagnosen. Systemet ble imidlertid godt motatt av legene, som glade og fornøyde tastet inn temperatur og puls for alle pasientene.

En mulig fremgangsmåte for å anvende dette er da å operere med forskjellige statistiske prediksjonsregler og la klinikerne velge hvilken de vil bruke, eller la klinikerne velge hvilke variabler prediksjonsregelen skal bruke i en liste av variabler med demonstrert prediktiv validitet. Selv om bare en av formlene kan være "best", er det sannynsligvis en rekke mulige formler som er bedre enn klinisk skjønn i de fleste tilfeller. Om en slik valgfrihet dermed øker bruken av beslutningsstøtte vil dette være av det gode, selv om en ikke "tvinger" klinikerne til å bruke den beste regelen til enhver tid. Dette forutsetter at man ikke anser det som ønskelig å "lure" klinikerne til å tro at de medvirker slik som beskrevet av Kaplan et al. og Berg. Dette er selvsagt også en mulighet, men bør kanskje holdes i bakhånd til man har prøvd andre metoder.

6.1.3 Konsekvenser av Yates et als funn

Yates et al (2003) kommer på bakgrunn av sine funn også med en del råd til utviklere av beslutningsstøtte. Vi husker at Yates et al. presenterte affektive reaksjoner og mangel på kausale forklaringer som mulige årsaker til at statistisk prediksjon ikke blir brukt. De lanserer to forslag for å motvirke dette.

Motvirk negativ affekt: Gå eksplisitt gjennnom og argumenter mot årsaker til negativ affekt. For å motvirke effekten av at en SPR ikke tar henyn til unike aspekter ved hver vurdering, kan man bruke likhetstanken for det den er verdt, og fremheve at alle behandles likt og rettferdig av formelen.

Kombiner skjønn og statistikk: Det bør være mulig å selge inn statistisk prediksjon hvis man lager beslutningsstøtteverktøyet slik at det ikke erstatter, men kommer i *tillegg* til klinisk skjønn. Den helhetlige kliniske vurderingen kan for eksempel brukes som en egen variabel i prediksjonsformelen. Er man litt kynisk kan man legge til at denne variabelen ikke trenger å vektes spesielt tungt, uten at man trenger å si dette til klinikerne.

Fremhev en vinner: På grunnlag av sine funn om at for mange muligheter virker overveldende anbefaler Yates at beslutningsstøtten bør kåre en klar "vinner", for eksempel ved å klart fremheve den mest sannsynlige blant en lang liste med differensialdiagnoser.

6.2 Opplysning

Meehl (1986) lanserte uvitenhet om forskningen som sentral årsak til manglende praktisk anvendelse, og dette beklages også blant annet av Arkes (2003), som kommenterer en artikkel i Science fra 1986 der noen geologer presenterte sin teori om at ekspertvurderinger kan kombineres. Hans kommentar er:

> "To scientists in the area of judgement and decision making, this article was not particurlarly newsworthy, because such methods of combing expert judgements have been in use for many years."
>
> — *Arkes, 2003; s. 5*

Økt fokus på artikler som Dawes et al. i Science (Dawes et al., 1989) og Swets et al. i Scientific American (Swets et al., 2000) er generelle virkemidler for opplysning av andre forskningsfelt. I Norge har Geir Kirkebøen en serie gående i A-magasinet angående bedømming og beslutningspsykologi generelt. Vi kan bare håpe at alle artiklene har en virkning, og at kunnskap om forskningen gradvis spres.

6.2.1 Hvordan presentere beslutningsstøtte for klinikere?

Som et underpunkt under "opplysning" inngår spesifikke metoder for å påvirke klinikernes holdninger til beslutningsstøtte. Bedømming og beslutningspsykologien er full av eksempler på faktorer som påvirker en vurdering, og det bør derfor være mulig å anvende noen av disse for å påvirke vurderingen om nettopp det å ta beslutningsstøtte i bruk. Yates et al. (2003) presenterer, i tillegg til de spesifikke rådene ovenfor, en del generelle råd for påvirkning av klinikerne.

Knus illusjonene: Bryt ned illusjonene om at intuisjon og klinisk skjønn er gode prediksjonsverktøy. Yates et al. oppfordrer til å gjøre dette ved hjelp av beslutningslogger for å redusere etterpåklokskap og kalibrere for overkonfidens. La klinikerne *oppleve* sine egne prediktive ferdigheter med og uten beslutningsstøtte, ikke bare *fortell* dem det.

Tapsaversjon: Tapsaversjon er en av de klassiske funnene fra bedømming og beslutningspsykologi (Kahneman and Tversky, 1979). Mennesker er mer sensitive for tap enn for vinning. I kombinasjon med framing (se nedenfor) betyr dette at vi bør fokusere på unngåelse av feil heller enn økning i korrekte prediksjoner, selv om dette er to sider av samme sak.

Framing: Når man så tar tapsaversjon med i betraktningen kan man benytte framing. Framing betyr bare at man presenterer de samme fakta fra et annet perspektiv, og dette kan enkelt gjøres i forhold til statistisk prediksjon. Har man en statistisk prediksjonsregel som har en PPV på 70%, i en setting der klinisk skjønn har en PPV på 60%, så presenterer vi det ikke som "absolutt økning i treffsikkerhet med 10%", eller "absolutt reduksjon av feil fra 40% til 30%", men som "25% reduksjon av feil!". Om man på denne måten fokuserer på den relative fordelen og samtidig på unngåelse av negative utfall kan man faktisk nyttiggjøre seg tapsaversjon, og gi klinikerne insentiv til å mene at beslutningsstøtte bidrar til en "god" prosess.

Fokuser på besparelser: En kan også nyttiggjøre seg tapsaversjon ved å fokusere på tidsbesparelsen som ofte er et resultat av beslutningsstøtte. Klinikerne sparer tid de kan bruke på andre ting, og organisasjonene sparer ressurser totalt sett. Gevinster på lang sikt bør spesielt fremheves da ett kritere for en "god" prosess var det å tendere mot å være best på sikt.

Minimer belastning: Yates et al. fant at beslutninger som innbærer mye arbeid ses på som vanskelige. Hvis vi da fokuserer på beslutningsstøtte som tids og arbeidsbesparende, og aller helst også klarer å gjøre beslutningsstøtten interessant og morsom kan dette bidra til at beslutningsstøtte ses på som en måte å gjøre vanskelige beslutninger lettere på.

6.3 Overtalelse

Parallelt med alt annet har vi sett at litteraturen på dette området ikke bare bærer preg av opplysning, men også av overtalelse. Om man kjenner til alle faktaene, men allikevel ikke vil bruke beslutningsstøtte av mer vage årsaker, som Yates et als "dårlige prosess", kan det kanskje hjelpe med direkte overtalelse. Den som har stått i fremste linje på dette feltet er uten tvil Dawes (2002a, 2005) med sine etiske argumenter. Hvorvidt dette har noen effekt er vanskelig å si. Jeg har ikke funnet noen tilsvar til noen av Dawes artikler, de eneste som refererer noen av disse etiske artiklene er faktisk Ægisdottir et al. (2006). Ytterligere forskning bør altså til for å undersøke hvilken effekt (om noen) overtalelse har på bruk av beslutningsstøtte.

6.4 Integrering i klinikk

Det synes rimelig å anta at en del av klinikeres holdninger til beslutningstaking formes i løpet av studiet. Om man ikke klarer å overtale dem til å bruke beslutningsstøtte etter studiet kan man kanskje instruere dem i å bruke det *under* studiet? Å få statistiske prediksjonsregler innarbeidet i studieløpet, også i klinikk, bør kunne påvirke bruksmønstre også etter studietiden. Sieck og Arkes (2005) fant jo også at generell holdning til statistikk predikerte bruk av beslutningsstøtte. Fokus på positiv holdning til statitsikk, og integrering av statistisk metode gjennom hele studieløpet bør derfor etterstrebes.

Om man på denne måten klarer å endre nyutdannede helsearbeideres holdninger til beslutningsstøtte og statistikk generelt vil man på lang sikt kunne

øke bruk av beslutningsstøtte i klinisk praksis. De eksisterende klinikerne lar seg imidlertid neppe påvirke så lett. Andre virkemidler bør derfor vurderes.

6.5 Tvang

Som vist er det massiv evidens for at statistisk prediksjon kan forbedre beslutninger innenfor helsesektoren (Ægisdottir et al., 2006), og også evidens for at mottakelsen av slik beslutningsstøtte er betydelig mindre positiv enn man kunne ønsket (Arkes, 2003). Det er også demonstrert massiv underbruk selv der statistiske prediksjonsregler er tilgjengelige (Corey & Merenstein, 1987; i Arkes, 2007). Med tanke på de konsekvenser dette har for pasienter og ressursbruk bør det derfor være i samfunnets interesse å vurdere andre tiltak, som bruk av tvang. Kliniske helsearbeidere er allerede underlagt en rekke lover og regler, og et ikke ubetydelig antall interne prosedyrer skal følges. Pålagt bruk av statistiske prediksjonsregler der de er tilgjengelige bør derfor være en mulighet. Om man etter innføring og pålagt bruk ser en reduksjon av feildiagnostisering og/eller feilbehandling bør det være lettere å få klinikerne til å endre oppfatning om beslutningsstøtten. Dette kan også føre til at klinikere selv ønsker å utarbeide beslutningsstøttesystemer, for på den måten å ha noe kontroll, og slippe å få tredd "fremmede" prosedyrer ned over hodet. Et slikt sterkt signal fra ansvarlige myndigheter bør ha en tilleggseffekt i at det henleder klinikernes oppmerksomhet mot forskningen, og kanskje dermed reduserer overkonfidensen om intuitive prediktive evner.

På den annen side er helsearbeidernes interesseorganisasjoner sterke pressgrupper i samfunnet, som det sjelden lønner seg å legge seg ut med. Rent politisk er derfor bruk av tvang svært risikabelt, i alle fall om man ikke først skaffer seg støtte hos inflytelsesrike personer både innen politisk ledelse og i de respektive interesseorganisasjoner. Får man først debatten i gang synes det imidlertid å være vanskelig å argumentere mot mer utstrakt bruk av klinisk beslutningsstøtte, både fra et vitenskapelig og etisk perspektiv.

Som et noe mindre drastisk virkemiddel kan det vurderes å innføre insentiver for økt utarbeidelse og bruk av statistiske prediksjonsregler, eller en kombinasjon av insentiver og tvang.

6.6 Alternativer til beslutningsstøtte: debiasing av klinisk skjønn

Et mulig alternativ til beslutningsstøtte ville jo være om det var mulig å forbedre klinisk skjønn, slik at forskjellene i treffsikkerhet utjevnes eller forsvinner. Som vi har sett er desverre skjevheter som overkonfidens svært vanskelige å få has på, og dette gjelder i større eller mindre grad de fleste slike mangler ved klinisk skjønn. Det er også vanskelig å se hvordan det er mulig å gjøre det bedre enn en statistisk prediksjonsregel som er utarbeidet gjennom en multippel regresjonsanalyse med mindre det eksisterer ukjente variabler med stor prediktiv kraft som er tilgjengelige for klinikeren, men ikke finnes i statistikken. En regresjonsanlayse vil jo ellers finne den optimale verdien til hver prediktor, og kan således brukes som en demonstrasjon av hvor bra det er mulig å gjøre det. Variasjon som ikke forklares av en regresjonsformel kan attribueres til sann usikkerhet. Det er imidlertid ikke utført voldsomt mye forskning på effekten av såkalt debiasing, og det er godt mulig det er mulig å forbedre klinisk skjønn i noen grad. En grundig diskusjon av dette ligger imidlertid utenfor skopet til denne oppgaven. For et overblikk over debiasing-metoder, se Larrick (2004).

6.7 Opplysning av pasienter

Et siste tiltak kan være å forsøke å motvirke (den hypotetiske) effekten av at klinikere tar hensyn til pasientenes holdninger til beslutningsstøtte. For å motvirke dette kan opplysning av pasienter være en mulig innfallsvinkel. Å endre holdninger er imidlertid ikke alltid like lett, og pasienters toleranse for informasjon de mener er uvesentlig er ikke nødvendigvis på topp, i alle fall ikke hvis de oppfatter at dette forsinker konkrete tiltak. Opplysning om forskningsresultatene innen beslutningsstøtte kan på sikt få store effekter. Hvis man klarer å få kunnskapen ut til "folk flest" er det gode muligheter for at det blir et folkekrav med økt bruk av beslutningsstøtte innen helsesektoren, i alle fall om man skal dømme etter medias ramaskrik hver gang det avdekkes en ørliten suboptimalitet i helsevesenet. Generell formidling av forskningen til befolkningen som helhet bør derfor etterstrebes.

Kapittel 7

Konklusjon

Beslutningsstøtte i form av statistisk prediksjon har gjennom 60 år demonstrert et potensiale for å forbedre prediktive vurderinger på en lang rekke områder, deriblant helsesektoren (Dawes et al., 1989; Grove et al., 2000; Ægisdottir et al., 2006). Økt bruk av statistisk prediksjon vil dermed kunne bidra til en betydelig reduksjon av feildiagnostiseringer og feilbehandlinger i helsesektoren (Hunt et al., 1998). På tross av dette har det vist seg svært vanskelig å innføre statistiske prediksjonsregler i praksis (Arkes, 2003; van Steenkiste et al., 2007). Årsakene til dette er mange, og kan blant annet spores til kunnskap om og faktorer ved beslutningsstøtte generelt (Meehl, 1986; Yates et al., 2003), beslutningstakeres forhold til teknologi (Grundmeier and Johnson, 1999; Naquin and Kurtzberg, 2004; Promberger and Baron, 2006), organisatoriske forhold (Arkes, 2003), lege-pasientforholdet (Pezzo and Pezzo, 2006; Arkes et al., 2007) og velkjente psykologiske faktorer som overkonfidens (Roese, 1997; Sieck and Arkes, 2005).

Mulige botemidler er mange, men forskningen på området er foreløpig begrenset. Noen få faktorer ved statistisk prediksjon er imidlertid påvist å ha noe effekt (Kaplan et al., 2001). Dette er i stor grad faktorer som manipulerer beslutningstakernes overkonfidens. Andre foreslåtte tiltak kan være å utnytte psykologiske faktorer som tapsaversjon og framing (Yates et al., 2003), generell opplysning om konsekvenser ved bruk og underbruk og overtalelse (Dawes, 2002; Dawes, 2005), integrering av beslutningsstøtte i profesjonsstudiene og mer drastiske tiltak som tvang og pålagt bruk ved å påvirke politiske myndigheter. En generell observasjon er at det må skilles mellom objektive resultater av beslutningsstøtteverktøy isolert sett, og mulighetene for at beslutningstakerne ønsker å benytte beslutningsstøtteverktøyene. Gode verktøy alene ser ikke ut til å være nok til å garantere bruk.

Forbedring av prediktive vurderinger i helsesektoren har altså vist seg svært vanskelig i gjennomføre i praksis, selv om teorien er sunn. Årsaker til og botemidler mot dette er enda ikke ikke godt nok forstått til å konkludere i den ene eller den annen retning. Videre forskning bør søke å klarlegge ytterligere faktorer som påvirker bruk av beslutningsstøtte i praksis, både i form av strukturelle faktorer ved beslutningsstøtten og psykologiske og organisasjonsmessige faktorer ved beslutningstakere og organisasjoner. Det er også utført skuffende lite forskning som undersøker effekten bruk av beslutningsstøtte har for pasientutfall, noe som forhåpentligvis får økt fokus i fremtiden.

Bibliografi

Arkes, H. R. (2003). The nonuse of psychological research at two federal agencies. *Psychological Science*, 14(1):1–6.

Arkes, H. R., Shaffer, V. A., and Dawes, R. M. (2006). Comparing holistic and disaggregated ratings in the evaluation of scientific presentations. *Journal of Behavioural Decision Making*, 19(5):429–439.

Arkes, H. R., Shaffer, V. A., and Medow, M. A. (2007). Patients derogate physicians who use a computer-assisted diagnostic aid. *Medical Decision Making*, 27(2):189–202.

Bar-Hillell, M. (1980). The base-rate fallacy in probability judgements. *Acta Psychologica*, 44:211–233.

Berg, M. (1997). *Rationalizing Medical Work*. MIT Press.

Copeland, J. (1993). *Artificial Intelligence - A Philosophical Introduction*. Blackwell.

Dawes, R. M. (2002). The ethics of using or not using statistical prediction rules in psychological practice and related consulting activities. *Philosophy of Science*, 69(3):178–184.

Dawes, R. M. (2005). The ethical implications of paul meehl's work on comparing clinical versus actuarial prediction methods. *Journal of Clincal Psychology*, 61(10):1245–1255.

Dawes, R. M., Faust, D., and Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674.

Grove, W. M. and Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public policy and Law*, 2(2):293–323.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., and Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assesment*, 12(1):19–30.

Grundmeier, R. and Johnson, K. (1999). Housestaff attitudes toward computer-based clinical decision support. In *Proceedings of the American Medical Informatics Associations Symposium, 1999*, pages 266–270. American Medical Informatics Associations.

Hunt, D. L., Haynes, R. B., Hanna, S. E., and Smith, K. (1998). Effects of computer-based clinical decision support systems on physician performance and patient outcomes - a systematic review. *JAMA - Journal of the American Medical Association*, 280(15):1339–1346.

Iyengar, S. S., Wells, R. E., and Schwartz, B. (2006). Doing better but feeling worse - looking for the 'best' job undermines satisfaction. *Psychological Science*, 17(2):143–150.

Kahneman, D. and Tversky, A. (1979). Prospect theory - analysis of decision under risk. *Econometria*, 47(2):263–291.

Kaplan, S. E., Reneau, J. H., and Whitecotton, S. (2001). The effects of predictive ability information, locus of control, and decision maker involvement on decision aid reliance. *Journal of Behavioral Decision Making*, 14(1):35–50.

Klein, G. (1999). *Sources of Power: how people make desicions*. MIT Press.

Larrick, R. P. (2004). Debiasing. In Koehler, D. J. and Harvey, N., editors, *Blackwell handbook of judgment and decision making.*, pages 316–337. Malden, MA: Blackwell.

Leger, C., Politis, D. N., and Romano, J. P. (1992). Bootstrap technology and applications. *Technometrics*, 34(4):378–398.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Asessment*, 50(3):370–375.

Naquin, C. E. and Kurtzberg, T. R. (2004). Human reactions to technological failure: How accidents rooted in technology vs. human error influence judgments of organizational accountability. *Organizational Behavior and Human Decision Processes*, 93(2):129–141.

Pezzo, M. V. and Pezzo, S. P. (2006). Physician evaluation after medical errors: Does having a computer decision aid help or hurt in hindsight? *Medical Decision Making*, 26(1):48–56.

Plous, S. (1993). *The Psychology of Judgement and Decision Making*. McGraw-Hill.

Promberger, M. and Baron, J. (2006). Do patients trust computers? *Journal of Behavioural Decision Making*, 19(5):455–468.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, 121(1):133–148.

Russell, E. W. (1995). The accuracy of automated and clinical detection of brain-damage and lateralization in neuropsychology. *Neuropsychology Review*, 5(1):1–68.

Sieck, W. R. and Arkes, H. R. (2005). The recalcitrance of overconfidence and its contribution to decision aid neglect. *Journal of Behavioural Decision Making*, 18(1):29–53.

Sim, I., Gorman, P., Greenes, R. A., Haynes, R. B., Kaplan, B., Lehmann, H., and Tang, P. C. (2001). Clinical decision support systems for the practice of evidence-based medicine. *Journal of the American Medical Informatics Association*, 8(6):527–534.

Sosial og helsedirektoratet (2006). Odelstingsproposisjon 49 (2005-2006) om lov om endringer i helseregisterloven (norsk pasientregister). Technical report, Sosial og helsedirektoratet.

Stang, E. (1996). Chernobyl - system accident or human error? *Radiation Protection Dosimetry*, 68(3/4):197–201.

Swets, J. A., Dawes, R. M., and Monahan, J. (2000). Better decisions through science. *Scientific American*, 283(4):82–87.

Teasdale, G. and Bennet, B. (1974). Assessment of coma and impaired consciousness - practical scale. *Lancet*, 2(7872):81–84.

van Steenkiste, B., van der Weijden, T., Stoffers, H. E. J. H., Kester, A. D. M., Timmermans, D. R. M., and Grol, R. (2007). Improving cardiovascular risk management: a randomized, controlled trial on the effect of a decision support tool for patients and physicians. *European Journal of Cardiovascular Prevention & Rehabilitation*, 14(1):44–50.

Waljee, J. F., Rogers, M. A. M., and Alderman, A. K. (2007). Decision aids and breast cancer: Do they influence choice for surgery and knowledge of treatment options? *Journal of Clinical Oncology*, 25(9):1067–1073.

Yates, F. J., Veinott, E. S., and Patalano, A. L. (2003). Hard decisions, bad decisions: On decision quality and decision aiding. In L., S. S. and J., S., editors, *Emerging Perspectives on judgement and decision research*, pages 13–63. New York: Cambridge University Press.

Ægisdottir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., N., N. C., Lampropoulos, G. K., Walker, B. S., Cohen, G., and Rush, J. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *Counseling Psychologist*, 34(3):341–382.