

Et vitenskapelig skjønn



Veiledere i psykisk helsevern anbefaler klinikere å bruke standardiserte vurderingsformer. Bruken av klinisk skjønn havner i skyggen. Men klinisk skjønn er også en vitenskapelig fremgangsmåte, noe Thomas Bayes underbygde allerede i 1763.

TEKST

Håkon Kongsrud Skard
Robert Ruts kij

PUBLISERT 1. mars 2017

I psykisk helsevern er det et økende fokus på standardisering av vurderinger ved førstegangskontakt med pasienter. Formålet er å redusere antall feilvurderinger samt fange opp høyrisikotilstander som psykose og suicidalfare (Helsedirektoratet, 2006; 2013). Standardiseringen innebærer at klinikeren bruker validerte symptom mål, eller stiller empirisk validerte spørsmål, som kan fange opp høyrisikotilstander – en form for «screening». Det som i liten grad vektlegges i veilederne og ved innføringen av slike rutiner, er betydningen av klinisk skjønn. Denne utelatelsen kan medføre flere feilvurderinger, stikk i strid med intensjonen bak standardiseringen. Effekten av screening som fenomen er heftig debattert ved ulike somatiske tilstander, som brystkreft og prostatakreft (Johansen, 2008), men innføringen av liknende prosedyrer i psykisk helsevern ser ikke ut til å ha utløst den samme debatten.

Et eksempel fra somatikken kan demonstrere risikoen ved å tillegge resultatet på standardiserte mål overdreven vekt, og illustrerer hvorfor innføringen av screening er omstridt. La oss si at du skal teste en pasient for en kreftform. Du vet at denne

kreftformen forekommer hos 1 % av befolkningen, og du vet at testen du bruker, er 99 % treffsikker. Testen utføres, og den er positiv. Hva er sannsynligheten for at pasienten har kreft? Mange ville intuitivt svart at de er 99 % sikre på at pasienten har kreft, men de som husker noe om baserateproblematikk fra statistikken, ville vært mer usikre. Svaret er at det er 50 % sannsynlighet for at pasienten har kreft. En test med 99 % treffsikkerhet er sjelden kost selv i somatikken og i hvert fall i psykisk helsevern. La oss si at treffsikkerheten er 90 %, som fortsatt er veldig bra. Med denne treffsikkerheten vil sannsynligheten for å ha kreft i det foregående eksempelet være på under 50 % *selv ved to positive tester* (Horgan, 2016). Det er i disse situasjonene det er lett å gjøre feilslutninger. Hvis en pasient har tatt to tester som du har stor tillit til, og begge viser samme resultat, hva er din naturlige reaksjon? Konsekvensene for pasienten vil være dramatiske uansett: enten en påvist livstruende sykdom – eller en krevende og kostbar behandling for en sykdom som ikke er der.

Måten vi regner oss frem til sannsynligheten på i eksempelet over er gjennom Bayes teorem. Bayes teorem er en statistisk tilnærming der relevant kontekstuell informasjon (kreftformens prevalens og instrumentets sensitivitet) inkluderes i vurderinger av hvor sannsynlig det er at en gitt pasient har en lidelse. Vi vil hevde at psykologer og psykiateres kliniske skjønn er grunnleggende likt denne tilnærmingen, og at betydningen av skjønn bør vektlegges i mye større grad når helsemyndigheter rådgir om vurderingssituasjoner.

Formålet med denne teksten er todelt. I første del vil vi vise hvordan klinikere er vitenskapelige når de gjør skjønnsmessige vurderinger av pasienter. I andre del vil vi argumentere for at slike skjønnsmessige vurderinger fordrer klinisk ekspertise. Et ensidig fokus på standardiserte metoder kan være uheldig, særlig for nyutdannede klinikere.

Bayesiansk logikk

Bayesiansk logikk er like vitenskapelig som den dominerende statistiske tankegangen bak standardiserte mål, men baserer seg på andre fundamentale prinsipper. Essensen i bayesiansk resonnering er å starte med en forutgående (a priori) sannsynlighet og ta inn ny informasjon for deretter å danne en etterfølgende (a posteriori) sannsynlighet. Dette kan uttrykkes matematisk med Bayes teorem (Kirkwood & Sterne, 2003):

$$P(B \text{ gitt } A) = P(B) * P(A \text{ gitt } B) / P(A)^{[1]}$$

Brakdal (2015) påpeker at erfarne klinikere er «bayesianere», uten at de er klar over at de benytter denne metoden, fordi de gjør det implisitt som en del av den diagnostiske prosessen. Innenfor evidensbasert medisin forutsettes det at klinikere er intuitive bayesianske tenkere (Guyatt, Rennie, Meade & Cook, 2008), og undersøkelser har vist at en bayesiansk tilnærming gir best diagnostisk presisjon ved mange somatiske tilstander, for eksempel diagnostisering av dyp venetrombose (Palareti, Cosmi & Legani, 2006) og akutt appendisitt (Sakai, Kobayashi, Nakamura, Toyabe & Akazawa, 2007).

Ved hjelp av noen fiktive kliniske eksempler ønsker vi å illustrere logikken i bayesiansk tenkning og vise hvor viktig klinisk skjønn / bayesiansk tenkning kan være i pasientvurderinger.

Tre kasus

Mange institusjoner i psykisk helsevern tilbyr øyeblikkelig hjelp, et tilbud der pasienter kan oppsøke bistand uten henvisning. Dette er situasjoner der pasienter skal vurderes opp mot potensielt alvorlige problemstillinger basert på lite forutgående informasjon. Vi ønsker å illustrere hvordan svaret på spørsmålet «Hører du stemmer?» kan ha helt ulik verdi og gi helt ulik forståelse basert på konteksten det gis i, ved å bruke tre fiktive «kasusjournaler»:

Pasient 1: Mann på 25 år, uten fast bopel, hentet av politi etter å ha gått til angrep på en statue og skreket til forbipasserende. Fremstår ustelt og uten adekvate klær for årstiden. Nekter å håndhile eller gi blikkontakt. Virker redd, snakker usammenhengende uten forutgående spørsmål, har lang latenstid ved spørsmål og gestikulerer vilt.

På spørsmål om han hører stemmer, svarer han: NEI.

Pasient 2: Kvinne på 50 år, oppgir ingen tidligere psykisk lidelse, ingen psykisk lidelse i familien. Jobber som sekretær, gift i 25 år, to velfungerende barn. Kommer til samtale med ektemann, gir god formell og emosjonell kontakt. Oppgir nedstemthet knyttet til at barna har flyttet ut. Ingen latenstid, svarer adekvat på spørsmål, motorisk rolig.

På spørsmål om hun hører stemmer, svarer hun at hun tror på ånder og kan av og til høre at de svarer. Så JA, hun hører stemmer.

Pasient 3: Ung velutdannet mann som kommer til samtale i følge med sin kjæreste. Bror med kjent psykoselidelse. Kjæresten har på forhånd ringt og varslet bekymring fordi hun opplever ham som underlig. Han har sagt opp jobben sin uten å informere henne, og hun har hørt ham snakke med seg selv. Velkledd og velstelt. Gir god formell kontakt. Viser klar motvilje mot undersøkelsen, bekrefter også dette selv. Mimikkfattig, svarer kortfattet og av og til etter lengre pauser. Når kjæresten ikke er med i samtalen, forteller han at han ikke føler han kan stole på henne lenger. Sier at han føler at han er i en livskrise, men ønsker ikke noe hjelp fra psykisk helsevern.

På spørsmål om han hører stemmer, svarer han: NEI.

Vi påstår at de fleste klinikere ville vurdert pasient 1 som sannsynligvis psykotisk, pasient 2 som sannsynligvis ikke psykotisk og pasient 3 som et tvilstilfelle, der sannsynligheten for psykose er usikker. Pasient 1 og 2 blir i så fall vurdert i strid med informasjonen innhentet fra det standardiserte spørsmålet, mens i pasient 3 sitt tilfelle så vil spørsmålet og pasientens svar vektlegges i vurderingen. At svaret vurderes ulikt, er fordi klinikere tar med all tilgjengelig informasjon i vurderingen og vektlegger svaret

på spørsmålet ut fra den forutgående informasjonen, i tråd med bayesiansk tenkning. La oss illustrere denne prosessen matematisk for å vise den underliggende logikken.

A priori-vurderinger

Forholdet mellom den diagnostiske testen, A (f.eks. symptom, laboratorieprøve, røntgenbilde, spørsmål), og sykdommen, B, uttrykkes statistisk gjennom Bayes teorem. I vårt tilfelle betyr dette at selvrapportert stemmehøring (diagnostisk test A +/-) og psykose (sykdom B +/-)^[2] skal opptre samtidig.

Initialt anslås en a priori-sannsynlighet for sykdommen (psykose), basert på den informasjonen behandleren besitter før testen. En a priori-vurdering omtales i litteraturen ofte som indikasjon, prevalens eller pretestsannsynlighet. Vi vil i det følgende kun bruke begrepet prevalens.

Vi definerer tre prevalensgrupper:

1. Lavprevalensgruppe: Hvis en utelukker den informasjonen som fremkommer i vurderingen utover pasientenes svar, vil prevalensen være lik prevalensen av psykose i befolkningen (Kringlen, 2001). Dette gjelder for eksempel for «Kvinne på 50» (pasient 2) $P(B) = 0,2\%$.
2. Middelsprevalensgruppe: Når forutgående informasjon gir sterk tvil om hvorvidt pasienten er psykotisk, som for «Ung velutdannet mann» (pasient 3), kan prevalensen angis som 50 %. $P(B) = 50\%$.
3. Høyprevalensgruppe: Når forutgående informasjon entydig peker i retning av at pasienten hører stemmer, for eksempel som i tilfellet med «Mann på 25 år» (pasient 1), kan prevalensen angis å være 80 %. $P(B) = 80\%$.

Testen utføres (spørsmålet stilles), og det aktuelle resultatet av en test betraktes på ethvert tidspunkt som en betinget sannsynlighet, det vil si at sannsynligheten for at resultatet er korrekt, tolkes ut fra forutgående informasjon. Dette gir grunnlaget for å gjennomføre følgende utregninger:

Sannsynlighet for å være psykotisk når personen bekrefter stemmehøring

$P(B+ \text{ gitt } A+)$ kalles for 'positiv prediksjonsverdi' og viser sannsynligheten for å være syk gitt en positiv test. I vårt tilfelle er positiv prediksjonsverdi sannsynligheten for å være psykotisk når personen bekrefter stemmehøring.

**«Innenfor evidensbasert
medisin forutsettes det at
klinikere er intuitive
bayesianske tenkere»**

$P(A+ \text{ gitt } B+)$ kalles for 'sensitivitet' og er lik sannsynligheten for en positiv test gitt at personen er syk. I vårt tilfelle: sannsynligheten for at en som er psykotisk (og hører

stemmer), bekrefter stemmehøring. Vi har ikke funnet data på hvor stor andel av de som faktisk er psykotiske, som vil bekrefte at de hører stemmer dersom de opplever dette. Derfor vil vi gjøre en skjønnsmessig vurdering av sensitivitet og bruke den videre i våre utregninger. Basert på denne vurderingen anslår vi at 90 % av psykotiske pasienter ville svare ja på et spørsmål om stemmehøring, $P(A+ \text{ gitt } B+) = 90\%$.^[3] $P(A)$ settes lik forekomsten av selvrapportert stemmehøring i befolkningspopulasjonen. Denne er forsiktig estimert til 4 % (Beaven et al., 2011).

Sannsynlighet for å ikke være psykotisk når personen benekter stemmehøring

Tilsvarende resonnering kan gjøres for negative funn. $P(B- \text{ gitt } A-)$ kalles for 'negativ prediksjonsverdi' og er lik sannsynligheten for å være frisk gitt en negativ test. I vårt tilfelle sannsynligheten for å ikke være psykotisk dersom man benekter stemmehøring.

$P(A- \text{ gitt } B-)$ kalles for 'spesifisitet' og er lik sannsynligheten for en negativ test gitt at man er frisk. Med utgangspunkt i rapportert stemmehøring i normalbefolkningen er denne lik 96 % (Beaven et al., 2011).

Vi kan nå beregne positiv og negativ prediktiv verdi for lav-, middels- og høyprevalensgruppene. Dette vil vi illustrere gjennom bruk av en CHI-kvadrat-tabell (se figur 1).

Figur 1

TEST	SYK	FRISK	
+	SP (sann positiv) a	FP (falsk positiv) b	a+b
-	FN (falsk negativ) c	SN (sann negativ) d	c+d
	a+c	b+d	a+b+c+d

$$\text{Sensitivitet} = a/a+c = 0,9$$

$$\text{Spesifisitet} = d/b+d = 0,96$$

$$\text{Positiv prediktiv verdi (PPV)} = a/a+b$$

$$\text{Negativ prediktiv verdi (NPV)} = d/c+d$$

Resultater

Hva er da verdien av spørsmålet stilt til pasienten, basert på disse utregningene? Vil svaret påvirke vurderingen av om pasienten er psykotisk eller ikke? I det følgende synliggjør vi med tall hvordan a priori-vurderinger (prevalensvurderinger) påvirker fortolkningen av svaret på om pasienten hører stemmer.

Figur 2

	Positiv prediktiv verdi (%)	Negativ prediktiv verdi (%)
Lavprevalensgruppe	4,3	99,9
Middelsprevalensgruppe	96,0	91,0
Høyprevalensgruppe	99,0	70,5

Lavprevalensgruppe

Det er 4,3 % sannsynlighet for at en person fra lavprevalensgruppen (her: «Kvinne på 50») som svarer bekreftende på spørsmål om stemmehøring, er psykotisk. Det er 99,9 % sannsynlighet for at en person fra lavprevalensgruppen som svarer avkrefteende på spørsmål om stemmehøring, ikke er psykotisk.

Middelsprevalensgruppe

Det er 96 % sannsynlighet for at en person fra middelsprevalensgruppen (her: «Ung velutdannet mann») som svarer bekreftende på spørsmål om stemmehøring, er psykotisk. Det er 91 % sannsynlighet for at en person fra middelsprevalensgruppen som svarer avkrefteende på spørsmål om stemmehøring, ikke er psykotisk.

Høyprevalensgruppe

Det er 99 % sannsynlighet for at en person fra høyprevalensgruppen (her: «Mann på 25 år») som svarer bekreftende på spørsmål om stemmehøring, er psykotisk. Det er 70,5 % sannsynlighet for at en person fra høyprevalensgruppen som svarer avkrefteende på spørsmål om stemmehøring, ikke er psykotisk.

**«Vi har personlig sett
situasjoner der diagnoser har
blitt satt med henvisning kun til
utslag på et standardisert
instrument som begrunnelse»**

Utregningene bekrefter på denne måten den skjønnsmessige konklusjonen til klinikerens, nemlig at svaret til pasient 2 («Kvinne på 50 år») ikke ville påvirket vurderingen, mens svaret til pasient 3 («Ung velutdannet mann») var viktig for klinikerens vurdering. Dette fordi det erfarne kliniske skjønn er intuitivt bayesiansk. Sannsynligheten for å ikke være psykotisk (NPV) basert på svaret til pasient 1 («Mann på 25 år») viser det kraftige utslaget som gis av høy prevalens kombinert med vårt høye anslag på sensitivitet. Den erfarne klinikerens ville sannsynligvis likevel vektet svaret mindre enn den observerte atferden i sin vurdering, samt neppe slått seg til ro med et resultat som ga 30 % feilmargin.

Skjønn versus standardisering

Helhetlige vurderinger med kontekstuell informasjon har i våre fiktive eksempler og i den kliniske virkeligheten bedre prediktiv verdi. Dette vil ikke overraske erfarne klinikere, men formidles det til nyutdannede klinikere? «Helhetlig vurdering» er en formulering som går igjen i alt fra lov om psykisk helsevern til interne kvalitetssikringsrutiner, og det er ingen grunn til å betvile at bakgrunn, aktuell situasjon og status presens tas med i samtlige klinikers vurderinger. Klarer likevel alle å innarbeide og vektlegge informasjon fra instrumenter riktig i en helhetlig vurdering?

Vi mener at standardiserte instrumenter forblir den beste undersøkelsesmåten for å fullt ut utforske et fenomen, gitt at de har høy reliabilitet og validitet. Vi er også bevisste at instrumenter inneholder mer enn ett spørsmål, slik vi har benyttet i utregningene, samt at manualene til instrumentene spesifiserer at de skal være en del av en bredere utredning. Vårt poeng er at standardiserte spørsmål brukt feil kan føre til falske positive eller falske negative svar, noe som er en potensiell trussel mot pasientsikkerheten i form av overbehandling eller feilbehandling, som illustrert gjennom debatten rundt screeningprosedyrer i somatikken. Dette gjelder spesielt lavprevalente tilstander som psykose. Gjennom oppfordringene fra helsemyndigheter og helseforetak til å bruke standardisering for å minimere kilder til feilslutninger, gir det opphav til nye. Vi har personlig sett situasjoner der diagnoser har blitt satt med henvisning kun til utslag på et standardisert instrument som begrunnelse, samt pasienter som utredes eller henvises igjen og igjen grunnet et tvilsomt utslag på et standardisert mål tilbake i tid. Dette kan tyde på at en helhetlig vurdering ikke alltid forekommer.

Vi opplever likevel at det hovedsakelig gjøres gode vurderinger i norsk psykisk helsevern, fordi det er erfarne klinikere som gjør dem og dermed er i stand til å avgjøre om standardiserte spørsmål skal benyttes. Vi er derfor kritiske til pålagte standardiserte spørsmål som et forbedringstiltak i kvalitetssikringsarbeid. De problemene vi viser som kan oppstå ved feil bruk av disse spørsmålene, har potensial til å utgjøre en fare for pasientsikkerheten, særskilt i situasjoner med uerfarent personell fremfor erfarne klinikere. Det er dokumentert at erfarne klinikere gjør bedre kasusformuleringer enn uerfarne (Eells, Lombart, Kendjelic, Turner & Lucas, 2005). Den erfarne kliniker vil kunne se bort fra svaret på det standardiserte spørsmålet som en del av sin helhetsvurdering, mens en uerfaren kliniker lettere vil kunne feste for stor lit til de standardiserte instrumentene, som kan skilte med en lang forskningstradisjon og anbefalt bruk fra myndighetene. Grunnen til at erfarne klinikere gjør bedre vurderinger, er kanskje at de besitter et mye større tilfang av a priori-kunnskap og derfor utviser et bedre klinisk skjønn? Samtidig vil noen hevde at klinikere med lang fartstid potensielt har en tendens til å overvurdere prevalens, da de har vært utsatt for et lite representativt utvalg i sitt virke.

Konklusjon

Det er god grunn til at standardisering og empirisk validering innenfor psykisk helsevern har blitt mer vanlig i de senere tiår. Den store variabiliteten i tilbudet til

pasienter, dokumentert blant annet i Riksrevisjonens gjennomgang av psykisk helsevern (Dokument 3:5, 2002–2003), synliggjorde behovet for å styrke pasientenes rett til et noenlunde likeverdig tilbud og etterprøvbare diagnostiske vurderinger. Videre har studier av beslutningsprosesser vist at vår intuisjon og våre iboende kognitive slagsider lett leder oss til feilslutninger (Kahnemann, 2011). Likevel, når Helsedirektoratets anbefalinger skal oversettes til implementering av helseforetakene, fører en prioritering av standardisering til en nedprioritering av klinisk skjønn og erfaring. Vi har vist at dette kan ha problematiske konsekvenser, da standardiserte mål kun er én av mange viktige kilder til vesentlige data. Denne problemstillingen er særskilt aktuell i psykisk helsevern, da fenomenene man søker å avdekke, ikke er direkte verifiserbare. Det er prisverdig at det søkes å redusere antall «falske negative» i form av alvorlig psykisk lidelse som forblir uoppdaget, men det er påfallende hvor lite det problematiseres at dette øker sannsynligheten for «falske positive» med negative konsekvenser i form av feilbehandling, unødvendig bruk av ressurser og potensielt store konsekvenser for enkeltindividet. Vi mener en større vektlegging av klinisk skjønn og en økt bevissthet hos klinikere om at de er «intuitive bayesianere», vil kunne redusere disse negative konsekvensene. Balansegangen mellom klinisk skjønn og standardiserte vurderingsrutiner bør være gjenstand for en større debatt enn det har vært hittil.

Teksten sto på trykk første gang i Tidsskrift for Norsk psykologforening, Vol 55, nummer 3, 2017, side 294-300

TEKST

Håkon Kongsrud Skard, psykolog, Diakonhjemmet sykehus

Robert Rutsikj, overlege, Ahus sykehus

+ [Vis referanser](#)

Referanser

Bevan, V., Read, J. & Cartwright, C. (2011). The prevalence of voice-hearers in the general population: A literature review. (3), 281–292.

Brakedal, B. (2015). En kliniker og en bayesianer [kronikk]. , 1468–70.

Dawson, M., Youngquist, S., Bledsoe, J., Madsen, T., Bossart, P., Davis, V. & Barton E. (2010). Low-risk young adult patients with chest pain may not benefit from routine cardiac stress testing: a Bayesian analysis. (3), 170–173.

Dokument 3:5 (2002–2003). Riksrevisjonens undersøkelse av psykisk helsevern - opptrappingsplanen 1999–2006.

Eells, T.D., Lombart, K.G., Kendjelic, E.M., Turner, L.C. & Lucas, C.P. (2005). The quality of psychotherapy case formulations: A comparison of expert, experienced and novice cognitive-behavioural and psychodynamic therapists. (4), 579–589.

Herrle, S.R., Corbett, E.C. Jr., Fagen, M.J., Moore, G.C. & Elnicki, D.M. (2011). Bayes' theorem and the physical examination: probability assessment and diagnostic decision making.

(5), 618-627, 2011.

Horgan, J. (2016). Bayes Theorem: What's the big deal? Hentet 12. mars 2016 fra <http://blogs.scientificamerican.com/cross-check/bayes-s-theorem-what-s-the-big-deal/>

Jackson, B.R. (2008). The dangers of false-positive and false-negative test results: false-positive results as a function of pretest probability. (2), 305-319.

Johansen, T.E.B. (2008). PSA-basert screening for prostatakreft. (22), 2612-2614.

Kringlen, E., Torgersen, S. & Cramer, V. (2001). A Norwegian Psychiatric epidemiological Study. (7), 1091-1098.

Leeflang, M.M., Bossuyt, P.M. & Irwig, L. (2009). Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. (1), 5-12.

Nichelatti, M. & Montemoli, C. (2008a). Bayesian statistics in medicine. Part I: The basic tools. (3), 342-346.

Nichelatti, M. & Montemoli, C. (2008b). Bayesian statistics in medicine. Part II: main applications and inference. (4), 422-431.

Nouraei, S.A., Huys, Q.J., Chatrath, P., Powles, J. & Harcourt, J.P. (2007). Screening patients with sensorineural hearing loss for vestibular schwannoma using a Bayesian classifier. (4), 248-254.

Ohayon, M.M. (2000). Prevalence of hallucinations and their pathological associations in the general population. (2-3) 153-164.

Palareti, G., Cosmi, B. & Legnani C. (2006) Diagnosis of deep vein thrombosis. (7), 659-672.

Rosenhan, D.L. (1973). On being sane in insane places. , 250-258.

Sakai, S., Kobayashi, K., Nakamura, J., Toyabe, S. & Akazawa, K. (2007). Accuracy in the diagnostic prediction of acute appendicitis based on the Bayesian network model. (6), 723-726.

Van Stralen, K.J., Stel, V.S., Reitsma, J.B., Dekker, F.W., Zoccali, C. & Jager, K.J. (2009). Diagnostic methods I: sensitivity, specificity, and other measures of accuracy. (12), 1257-1263.

Veileder psykisk helsevern for voksne (2006) . Oslo: Helsedirektoratet. IS-1388.

Tilgjengelig fra:
<https://helsedirektoratet.no/Lists/Publikasjoner/Attachments/281/Distriktopsykiatriske-sentre-IS-1388.pdf>

Utredning, behandling og oppfølging av personer med psykoselidelser (2013). . Oslo:

Helsedirektoratet. IS-1957. Tilgjengelig fra:
<https://helsedirektoratet.no/Lists/Publikasjoner/Attachments/326/Nasjonal-faglig-retningslinje-for-utredning-behandling-og-oppfolging-av-personer-med-psykoselidelser-IS-1957.pdf>

www.scid4.org (2013). <http://www.scid4.org/faq/scidfaq.html>