Bilaga 6 Grundläggande terminologi för använda effektmått och översiktlig förklaring av genomförda analyser

Deskriptiv statistik för diagnostisk tillförlitlighet

De två helt centrala begreppen är sensitivitet (metodens känslighet att identifiera sjukdom) och specificitet (metodens träffsäkerhet att utesluta sjukdom). Sensitivitet definieras matematiskt som andelen sant positiva dividerat med summan för de sant positiva och falskt negativa svaren i diagnostiken som utvärderas. Specificiteten i sin tur definieras som de sant negativa dividerat med summan för de sant negativa och de falskt positiva (se nedan). Kliniskt föreligger vanligen ett reciprokt förhållande mellan sensitivitet och specificitet, om det ena stiger sjunker det andra.

Traditionell diagnostisk fyr-fälts tabell		*Dikotom förutsägelse (diagnostik) av sjukdom*	
		Positiv diagnostik	*Negativ diagnostik*
Faktiskt sjukdom (facit)	*Positiv*	*Sant Positiv,* ***SP*** *(indextest positivt, överensstämmer m referenstest)*	*Falskt Negativ,* ***FN*** *(indextest negativt, överensstämmer ej m referenstest)*
	Negativ	*Falskt Positiv,* ***FP*** *(indextest positivt, överensstämmer ej m referenstest)*	*Sant Negativ,* ***SN*** *(indextest negativt, överensstämmer m referenstest)*

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

$$Specificitet = \frac{SN}{SN + FP}$$

För att kunna utföra de matematiska beräkningarna av sensitivitet och specificitet krävs en tillgänglig och övergripande diagnostisk sanning, som de aktuella diagnostiska algoritmerna jämförs med. Detta representeras av referenstestet Det aktuella referenstestet i det här projektet förekommer i två likvärdiga former, dels en mikroskopisk undersökning på cellnivå, även kallat patologisk anatomisk diagnos, PAD, som utförs av dedicerade specialister, dels sjukdomsfrihet över tid. En sådan vedertagen tidsperiod är 12 månader och har använts som minimitid i denna sammanställning. Det innebär att den identifierade förändringen inte utvecklades vid strukturerade kontroller under perioden, på ett sådant sätt att diagnostikerns övertygelse avseende förändringens godartade natur ändrades.

Andra ofta använda kliniska begrepp

Under- respektive överdiagnostik

Underdiagnostik kännetecknas av att analysen underskattar det sanna antalet som har sjukdomen. Detta ses vid en låg sensitivitet, alltså då antalet falskt negativa är för högt. Dess kliniska konsekvens innebär detta att man behandlar alltför få i förhållande till det sanna antalet med faktiskt sjukdom.

Å andra sidan är det utmärkande för överdiagnostik att man överskattar det faktiska antalet med sjukdomen. Det vill säga att antalet falskt positiva är högt vilket leder till en låg specificitet och kliniskt att alltför många genomgår terapi.

Positivt- respektive negativt prediktivt värde (PPV och NPV)

Dessa två begrepp används kliniskt vanligen främst i samband med överlämnandet av svaret från en diagnostisk undersökning och den påföljande diskussionen. Begreppen anger hur stor sannolikheten är att provsvaret förutsäger (predicerar) en sanning. Det positivt prediktiva värdet (PPV) anger således sannolikheten att ett positivt svar faktiskt är sant och patienten har sjukdomen. Det negativa prediktiva värdet (NPV) anger sannolikheten att patienten inte har sjukdomen vid ett negativt utfall av diagnostiken.

$$PPV = \frac{SP}{SP + FP}$$

$$NPV = \frac{SN}{SN + FN}$$

Sammanvägningen av sensitiviteten och specificiteten

Den grundläggande förutsättning för en kunna genomföra en sammanvägning är att studiernas population, genomförande och använda tröskelvärden mm är tillräckligt likartade.

Ett kopplat skogsdiagram görs för att få en visuell illustration av de enskilda studiernas faktiska värden för sensitivitet respektive specificitet och hur de förehåller sig till varandra för varje algoritm (uppdelat på pre- respektive postmenopausalt status). Detta kopplade skogsdiagram är inte någon sammanvägning av sensitivitet och specificitet. Inte heller har de enskilda studiernas faktiska sensitivitet och specificitet viktats utifrån studiernas storlek.

Sammanvägningen görs i en bivariat metaanalys. Den kallas Summations Receiver Operating Curve, SROC om man använt sig av en så kallad Fixed-Effects Model. Denna väljs om studierna är tillräckligt likartade enligt ovan. Om en däremot viss och begränsad heterogenitet (som dock tillåter sammanvägning) mellan studierna föreligger väljs en hierarkisk SROC (HSROC). Skillnaden mellan SROC och HSROC är att i den sistnämnda använder man sig av en Random Effects Model.
I föreliggande rapport har HSROC-metodiken använts.

Sammanvägningen sammanfattas och plottas som ett punktestimat. Förutom punktestimatet erhålles även en konfidensregion, som anges med en viss säkerhet, vanligen 95 procent. Konfidensregionen kan jämförs med konfidensintervallet i en univariat analys.
Om det föreligger överlappande konfidensregioner mellan två olika punktestimat saknas tillräcklig visuell evidens för att vi ska kunna säga att punktestimaten är åtskilda. Om konfidensregionerna inte överlappar alls, är det ett starkt visuellt stöd för att grupperna har olika diagnostisk prestanda.
Det inte meningsfyllt att numeriskt ange konfidensregionens totala area. Detta eftersom konfidensregionen kan ha olika utbredning i X- och Y-led samtidigt som arean av regionen förblir konstant. Detta är grunden till att man gör en visuell bedömning av grafen om eventuellt överlapp finns.
I den bivariata analysen erhålles även en 95 procent prediktionsregion. Denna region förutsäger (med 95 % säkerhet) regionen/ytan inom vilket punktestimatet kommer att vara i vid en eventuell framtida studie genomförd under liknande betingelser.

Jämförelser mellan de olika algoritmerna

För att jämföra de olika algoritmerna med syftet att rangordna dem sinsemellan har tre separata metoder använts. Metoderna är alla indirekt jämförande, dvs alla patienter som deltagit i studierna har inte blivit studerade med alla sex ingående algoritmer. Vi har tillåtit oss att göra på detta sätt då de ingående studiernas vetenskapliga tillförlitlighet har bedömts som Låg eller Måttlig vid granskningen av risken för systematisk snedvridning enligt GRADE med instrumentet Quadas-2. Samt att samtliga medtagna studier kunde inkluderas utifrån rapportens Frågeställning, PIRO och Avgränsningar.
De tre metoderna var:

- Den framräknade diagnostiska tillförlitligheterna för de enskilda algoritmerna jämfördes med varandra med hjälp av HSROC-metoden (Figur 6.13 och 6.26). Punktestimat med tillhörande konfidensregioner plottades i denna bivariata modell för samtliga algoritmer. Med undantag för LR2 användes frekventistisk GLMM-modell (Generalized Linear Mixed Model) för beräkningarna. För LR2, vars totala antal observationen var otillräckligt för GLMM, användes istället Baysiansk inferens för att få fram motsvarande resultat. Den Baysianska användningen av priorfördelningar för modellparametrar hjälper till att stabilisera beräkningarna när datamängden är liten. Resultatet påverkas av vilka priors som används. I det här fallet använde vi de standardpriors som finns i programvaran MetaBayes DTA.

- Vidare beräknades den Diagnostiska Odds Ratio (DOR, Tabell 6.7 och 6.14). Det definieras som odds för att testet är positivt hos dom med sjukdom delat med odds för att testet är positivt hos dom utan sjukdom enligt formeln:

  $$DOR = (sensitivitet\ x\ specificitet)/(1 - sensitivitet)x(1 - specificitet)$$

  DOR är ett sammanfattande mått, men det säger ingenting om balansen mellan sensitivitet och specificitet. Två tester med samma DOR kan ha mycket olika klinisk användbarhet beroende på var de ligger på ROC-kurvan. Ju högre DOR-värdet är desto bättre är diagnostikens särskiljande förmåga, värdet 1 innebär ingen särskiljande förmåga

- Det så kallade Youden's Index beräknades enligt nedan

  $$Youden's\ index = sensitivitet + specificitet - 1$$

  Värden över 0,5 anses som acceptabla för ett test och för ett idealiskt test med sensitivitet 1,0 och specificitet 1,0 blir värdet 1. Används ofta för att identifiera en optimal tröskel i ROC-analys, det vill säga den punkt där summan av Sensitivitet och Specificitet är som störst.

  Varken för DOR eller Youden's s test har några osäkerhetsmått beräknats.

För ytterligare information, se SBU:s handbok, Cochrane handbook for Systematic Reviews eller annan ämnesspecifik litteratur.
I denna rapport har vi använt oss av programmen RevMan 5.4.1, MetaDTA, Diagnostic Test Accuracy Meta-Analysis v2.1.5 och MetaBayes DTA v1.5.2