Bilaga 3 Diagnostisk förmåga

Diagnostisk förmåga

Sensitivitet och specificitet

Den diagnostiska förmågan kan utvärderas med sensitivitet och specificitet som beräknas från en fyrfältstabell (Tabell 1).

Tabell 1. *2x2-tabell som används för att utvärdera diagnostiska metoder.*

	Har sjukdomen	**Har inte sjukdomen**	**Summa**
Test positivt (tyder på att sjukdomen finns)	a	b	a+b
Test negativt (tyder på att sjukdomen inte finns)	c	d	c+d
Summa	a+c	b+d	a+b+c+d

Sensitivitet = a / (a+c) (antalet sant sjuka med positiv test dividerat med totala antalet sant sjuka).

Specificitet = d / (b+d) (antalet sant friska med negativ test dividerat med totala antalet sant friska).

Positivt prediktivt värde = a / (a+b) (andelen som har sjukdomen av alla med postivt test).

Falskt positva = b / (a+b) (antalet sant friska med positivt testdividerat med totala antalet sant friska).

Negativt prediktivt värde = d / (c+d) (andelen som inte har sjukdomen av alla med negativt test).

Falskt negativa = c / (c+d) (antalet sant sjuka med negativt test dividerat med totala antalet sant sjuka).

”Accuracy” (träffsäkerhet) = (a+d) / (a+b+c+d) (andelen korrekt klassificerade patienter).

Av ovanstående sex variabler har sensitivitet och specificitet fördelen att inte vara beroende av prevalensen av sjukdomen i den undersökta populationen och det är därför som värden på

sensitivitet och specificitet kan generaliseras och användas för att utvärdera diagnostiska metoder. Även falskt negativa och falskt positiva är oberoende av prevalensen men dessa termer används inte så mycket eftersom de är lika med (1 - sensitivitet) respektive (1 – specificitet).

ROC-kurvor (receiver operating characteristics)

Vissa diagnostiska test som t ex förekomsten av glukos i urinen vid misstanke om diabetes ger enbart ett svar i kategorierna positivt eller negativt test. Vid utredning pga misstanke om BPO får man i stället resultatet av testet i form av en kontinuerlig variabel, ju större prostata är eller ju lägre det maximala flödet är desto större sannolikhet är det att sjukdomen finns. Då utfallet av det diagnostiska testet är en kontinuerlig variabel kan gränsen för vad man uppfattar som positivt och negativt test väljas olika. Ändras gränsen så att sensitiviteten ökar får man ”betala” för denna förbättring genom att specificiteten minskar och vice versa. Om den diskriminativa gränsen successivt ändras från värdet på den lägsta observationen till den högsta ändras sensitiviteten från 0 till 1 samtidigt som specificiteten går från 1 till 0. De parvisa värdena av sensitivitet och specificitet kan ritas in i ett ROC-diagram (figur 1).

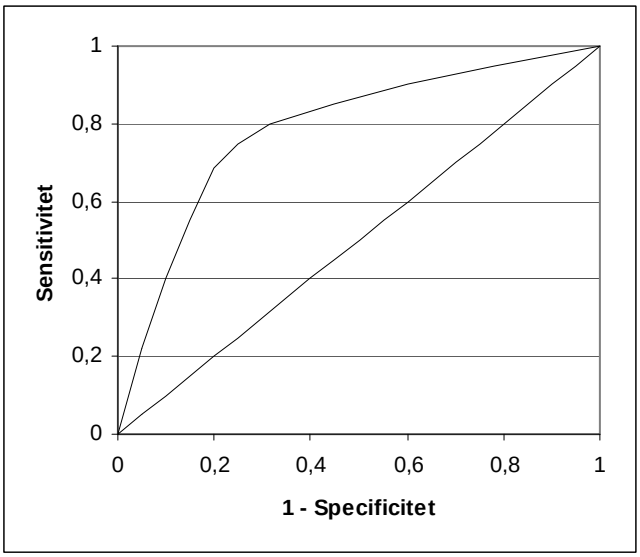


Fig 1. *ROC-diagram. Kurvan visar hur sensitivitet och specificitet ändras då det diskriminativa värdet successivt ändras från det mest patologiska värdet till det minst patologiska. Ju närmare det övre vänstra hörnet kurvan går desto bättre diagnostisk förmåga.*

Den diagonala linjen som går från punkten 0;0 till punkten 1;1 visar parvisa värden av sensitivitet och specificitet för en metod som inte har något diagnostiskt värde alls. En diagnostisk metod som fungerar perfekt och klassificerar alla patienter rätt har en kurva som börjar i punkten 0;0 och sedan följer y-axeln till punkten 0;1 varefter den går horisontellt till punkten 1;1. De flesta diagnostiska metoderna beskrivs av en kurva som ligger mellan dessa två extremer, ju längre upp mot det övre vänstra hörnet av diagrammet kurvan ligger desto bättre är den diagnostiska metoden.

En metod att kvantifiera den diagnostiska förmågan är att beräkna ytan under ROC-kurvan. För metoder utan diagnostiskt värde är ytan 0,5 och för perfekta metoder 1. En yta på 0,8 räknas som bra och 0,7 som måttligt bra. Ytan under kurvan är inte beroende av sjukdomens prevalens. En fördel med ROC-kurvor är att man kan jämföra den diagnostiska förmågan hos test utan att man har bestämt ett gränsvärde som skiljer sjuka och ej sjuka åt. Just inom LUTS och BPO är det ett problem att man använder olika diskriminativa värden.

Likelihood ratio (sannolikhetskvot)

Ett annat sätt att beräkna ett sammanfattande mått på den diagnostiska förmågan är att beräkna sannolikhetskvoter.

Positive likelihood ratio (LR+) = sannolikheten för ett positivt test om sjukdomen föreligger dividerad med sannolikheten för ett positivt test om sjukdomen ej föreligger = a / (a+c) / b / (b+d) = sensitivitet / (1 – specificitet).

Negative likelihood ratio (LR) = sannolikheten för ett negativt test om sjukdomen föreligger dividerad med sannolikheten för ett negativt test om sjukdomen ej föreligger = c / (a+c) / d / (b+d) = (1 – sensitivitet) / specificitet.

Likelihood ratios kan variera mellan 0 och oändligheten. Ett värde på 1 betyder att metoden inte har något diagnostiskt värde. Som positive likelihood ratio är definierad har den ett värde över 1 och ju högre värdet är desto bättre är den diagnostiska förmågan. Negative likelihood ratio är under 1 och ju närmare 0 värdet är desto bättre är den diagnostiska förmågan.

Likelihood ratios beräknas från sensitivitet och specificitet och de har alltså fördelen att inte vara beroende av sjukdomens prevalens. Vid kontinuerliga variabler är LR+ och LR– beroende av hur man sätter den diskriminativa gränsen för testet. Sätter man gränsen så att LR+ förbättras (ökar över 1) försämras samtidigt LR– (ökar under 1) och vice versa. Sätter man en snäv gräns för positivt test får man en låg sensitivitet men mycket hög LR+ eftersom

(nästan) bara sjuka individer har kraftigt patologiska värden. Omvänt får man med en vid gräns för positivt test en hög sensitivitet och en mycket låg (bra) LR– eftersom endast friska individer har ”supernormala” värden.

Ett visst värde för LR+ och motsvarande inverterade värde för LR– tyder på samma diagnostiska förmåga. Likelihood ratios på t ex 4 och 0,25 är alltså likvärdiga. Diagnostiska metoder som har ett likelihood ratio över 10 eller under 0,1 anses ha god diagnostisk förmåga.

När man jämför diagnostiska metoder måste både sensitivitet och specificitet respektive både LR+ och LR– vara bättre för den ena metoden för att man ska kunna säga att denna metod är bättre på att diagnostisera sjukdomen. I de fall en variabel är bättre och den andra sämre vet man inte säkert om metoderna är likvärdiga eller om den ena är bättre. Jämförelse av diagnostiska metoder är enkel om man väljer det diskriminativa värdet så att sensitivitet = specificitet. Då blir också automatiskt LR+ = 1 / LR–. Vid jämförelser blir det då också automatiskt så att både sensitivitet och specificitet respektive både LR+ och LR– blir bättre för den ena metoden om metoderna har olika diagnostisk förmåga.

Sannolikhet för sjukdom före respektive efter ett diagnostiskt test

Ett sätt att åskådliggöra nyttan av ett diagnostiskt test är att beskriva hur sannolikheten för sjukdom (och sannolikheten att ej ha sjukdomen) förändras när man utför testet. Sannolikheten att ha sjukdomen före testet är lika med prevalensen i den undersökta populationen. Prevalensen för BPO i en population av män med LUTS tydande på BPO och som sökt sjukvård ligger sannolikt kring 50 % och den bör för de flesta mottagningarna ligga mellan 25 och 75 %. Om prevalensen BPO är hög eller låg inom detta intervall bör i första hand bero på hur många patienter med lindriga miktionsbesvär och hur många män med centralnervösa åldersförändringar och cerebralt ohämmad blåsa som är inkluderade i den undersökta populationen.

Tabell 2 visar vilka LR+-värden som behövs för att öka sannolikheten för sjukdom från vissa sannolikheter före testet till vissa högre sannolikheter efter testet om det utfaller positivt. I tabell 3 visas hur man samtidigt minskar sannolikheten för sjukdom om testet utfaller negativt under förutsättning att LR– = 1 / LR+. LR+ och LR– bestäms av sensitivitet och specificitet. Gör man antagandet att sensitivitet = specificitet kan man räkna ut vilken sensitivitet/specificitet värdena på positive likelihood ratio i Tabell 3 motsvarar. Dessa värden på sensitivitet/specificitet visas i tabell 4. I de här tabellerna kan man se att diagnostiska metoder behöver vara bättre om sjukdomen är ovanlig och prevalensen är 0,01 än om

prevalensen är 0,50. Den diagnostiska förmågan skulle även behöva vara bättre i den orealistiska situationen att prevalensen av sjukdom var 0,99.

Tabell 2. *Tabellen visar vilka LR+-värden som behövs för att öka sannolikheten för sjukdom vid positivt test vid olika sjukdomsprevalenser.*

Sannolikhet efter test	**Sannolikhet för sjukdom före test = prevalens**						
	0,01	**0,10**	**0,25**	**0,50**	**0,75**	**0,90**	**0,99**
0,01	1						
0,05	5,2						
0,1	11	1					
0,25	33	3	1				
0,5	99	9	3	1			
0,75	297	27	9	3	1		
0,9	891	81	27	9	3	1	
0,95	1881	171	57,1	19	6,3	2,1	
0,99	9801	891	294	99	33	11	1

Tabell 3. *Tabell 2 visar hur sannolikheten för sjukdom ökar vid ett positivt test. Den här tabellen ska läsas tillsammans med tabell 2 och den visar hur sannolikheten för sjukdom minskar om testet istället är negativt under förutsättning att LR– = 1/LR+.*

	Sannolikhet för sjukdom före test = prevalens						
"Radnamn"	**0,01**	**0,10**	**0,25**	**0,50**	**0,75**	**0,90**	**0,99**
0,01	0,01						
0,05	0,0019						
0,1	0,0009	0,1					
0,25	0,0003	0,036	0,25				
0,5	0,0001	0,012	0,1	0,5			
0,75	<0,0001	0,0041	0,036	0,25	0,75		
0,9	<0,0001	0,0014	0,012	0,1	0,5	0,9	
0,95	<0,0001	0,0006	0,0058	0,05	0,32	0,81	
0,99	<0,0001	0,0001	0,0011	0,01	0,083	0,45	0,99

Tabell 4. *Tabellen visar vilken sensitivitet och specificitet LR-värdena i tabell 2 motsvarar under förutsättning att sensitivitet = specificitet.*

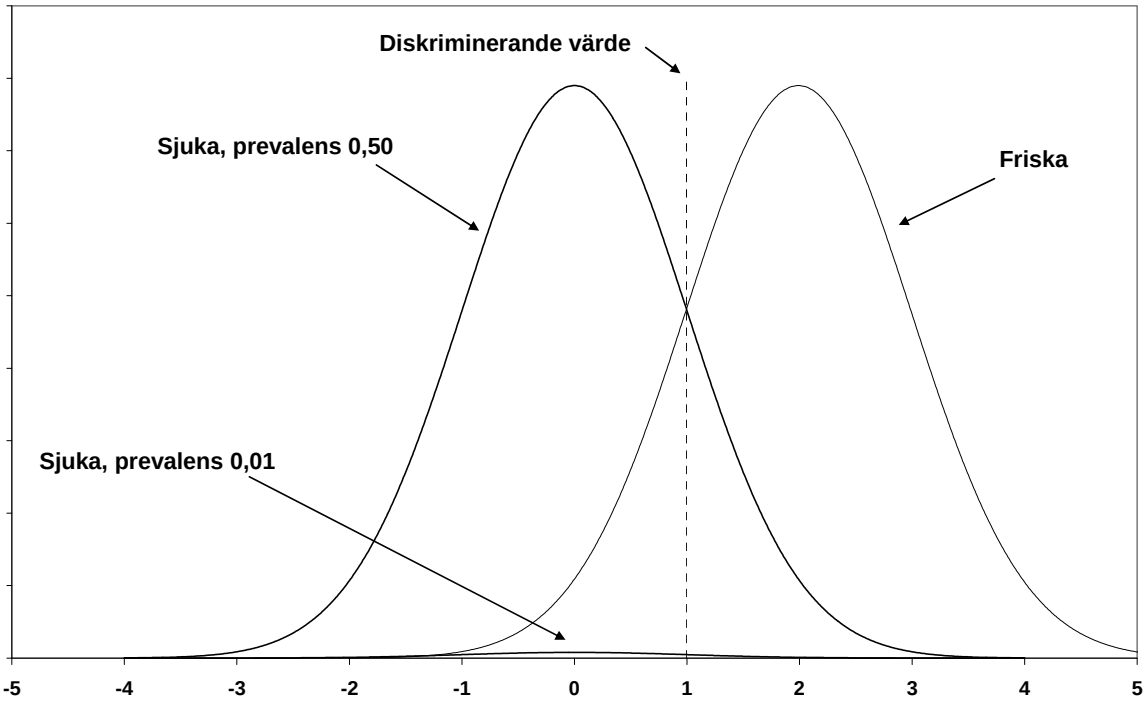
Sannolikhet efter test	Sannolikhet för sjukdom före test = prevalens						
	0,01	**0,1**	**0,25**	**0,5**	**0,75**	**0,9**	**0,99**
0,01	0,50						
0,05	0,84						
0,1	0,92	0,50					
0,25	0,97	0,75	0,5				
0,5	0,99	0,90	0,75	0,50			
0,75	0,997	0,96	0,90	0,75	0,50		
0,9	0,999	0,99	0,96	0,90	0,75	0,50	
0,95	>0,999	0,994	0,98	0,95	0,86	0,68	
0,99	>0,999	0,999	0,997	0,99	0,97	0,92	0,50

Ett annat sätt att visa att det behövs bättre diagnostisk förmåga vid ovanliga sjukdomar än vanliga är att se på accuracy, dvs andelarna korrekt och felaktigt klassificerade patienter. Vi antar först att risken för sjukdom är 0,01 och att vi använder ett måttligt bra diagnostiskt test med LR+ 5,2; LR– 1 / 5,2 och sensitivitet = specificitet 0,84. I detta fall har 84 % av patienterna klassificerats korrekt och 1 av 20 som klassificerat som sjuk har sjukdomen. Andelen falskt sjuka kan minskas genom att man ändrar den diskriminativa gränsen men man får en oacceptabelt låg sensitivitet innan man får en bra träffsäkerhet (accuracy). De här siffrorna ska jämföras med att man klassificerar 99 % korrekt om man säger att alla saknar sjukdomen utan att göra något test. För att använda ett sådant här test är det nödvändigt att sjukdomen som man diagnostiserar är betydelsefull och att det finns ett bra sekundärt test.

Situationen blir mycket annorlunda om ett test med denna diagnostiska förmåga används på en population där risken för sjukdom är 0,50. Med denna risk för sjukdom är det inte någon statistisk parameter som kan försämras utan även relativt dåliga diagnostiska test kan enbart förbättra klassificeringen. Om patienterna klassificeras med slumpen eller om man antar att alla har sjukdomen eller att alla är friska blir andelen felklassificerade 50 %. Genom att använda det diskuterade testet med sensitivitet och specificitet 0,84 minskas andelen felklassificerade till 16 %. I det här fallet är det relativt många patienter som är felaktigt klassificerade som sjuka eller friska. Om det är viktigt att undvika dessa felklassificeringar

behövs komplettering med ytterligare diagnostik. Om effekterna av en felklassificering är mindre, t ex att man ger eller inte ger symtomatisk medicinering, kan testet räcka till.

Skillnaderna mellan att diagnostisera en vanlig och ovanlig sjukdom framgår också av figur 2. För en ovanlig sjukdom kan det vara så att för i princip alla värden på den diagnostiska variabeln är det större sannolikhet att vara frisk än sjuk. Vid prevalensen 0,50 är det inte så utan på ena sidan av ett gränsvärde är det större sannolikhet att vara sjuk och på den andra att ej ha sjukdomen.



Figur 2. *Figuren illustrerar skillnaden på att diagnostisera tillstånd med hög resp låg prevalens. I båda fallen används ett diagnostiskt test med sensitivitet = specificitet = 0,84, LR+ = 5,2 och LR– = 1/5,2. Medelvärde och standarddeviation är identiska i de två sjukdomspopulationerna.*

Om man använder LR-kvoter kan man lätt se hur en serie av flera diagnostiska test påverkar sannolikheten för sjukdom. Antag att vi gör två diagnostiska test som har LR+ 3 och LR- 1 / 3. Om båda testen utfaller positivt ökar sannolikheten för sjukdom lika mycket som om vi gjort ett test med LR 3 *3, dvs 9. Om ett test utfaller positivt och ett negativt ändras sannolikheten som om man gjorde ett test med LR 3 * 1 / 3, dvs 1, vilket betyder att de diagnostiska testen inte har givit någon information. Om flera test med låg diagnostisk

förmåga är samstämmiga får man en bra diagnostisk säkerhet, om de ej är samstämmiga är diagnosen osäker. I det här sammanhanget måste man tänka på att likelihood ratio för ett test kan vara olika för patienter som inte gjort något test, för dem som gjort ett test med positivt utfall och för dem som gjort ett test med negativt resultat.

Korrelationer

Den fyrfältstabell som används för att beräkna sensitivitet och specificitet beskriver korrelationen mellan egenskapen att ha en sjukdom och utfallet av ett diagnostiskt test. De variabler som man använder i det diagnostiska testet och för att få fram vilka som är sant sjuka och sant friska kan antingen vara kvalitativa eller kontinuerliga. Vid kontinuerliga variabler måste man bestämma ett gränsvärde för att dikotomisera variablerna. I fallet med LUTS och BPO används diagnostiska metoder som ger utfallet i kontinuerliga variabler, t ex symtomskalor, prostatavolym, urinflöde, vilka ofta jämförs med ett kontinuerligt mått på uretraresistensen. Inom området LUTS och BPO finns det ett begränsat antal studier som redovisar sensitivitet, specificitet och LR-värden men det finns relativt många studier som visar korrelationen mellan olika variabler. Anledningen till detta är sannolikt dels att många studier är lite äldre och dels att det inte finns allmänt accepterade gränsvärden för de kontinuerliga variablerna.

Om man antar att både den diagnostiska variabeln och den som används för att definiera sjukdomen är normalfördelade kan man räkna ut vilken diagnostisk förmåga olika korrelationskoefficienter har. Även om detta antagande inte stämmer exakt får man sannolikt en relativt bra bild av den diagnostiska förmågan genom att bedöma korrelationskoefficienten. I Tabell 5 kan man se vilken sensitivitet och specificitet olika korrelationskoefficienter motsvarar och i Tabell 6 är dessa värden omräknade till LR+ och 1 / LR–.

I tabellerna kan man se att samma korrelationskoefficient har något bättre diagnostisk förmåga vid en ovanlig sjukdom, men sjukdomsprevalensen har inte så stor betydelse.

Korrelationskoefficienter från 0 till 0,5 motsvarar en svag diagnostisk förmåga, från 0,5 till 0,9 motsvarar en måttlig diagnostisk förmåga som räcker till om 2–3 test kombineras och koefficienter över 0,9 motsvarar bra diagnostisk förmåga.

Tabell 5. *Sensitivitet och specificitet för olika korrelationskoefficienter under förutsättning att sensitivitet = specificitet.*

Korrelations-koefficient	Sjukdomsprevalens				
	0,10	**0,30**	**0,50**	**0,70**	**0,90**
0	0,50	0,50	0,50	0,50	0,50
0,1	0,53	0,54	0,54	0,54	0,57
0,2	0,57	0,57	0,57	0,58	0,60
0,3	0,62	0,60	0,60	0,61	0,64
0,4	0,66	0,64	0,64	0,65	0,69
0,5	0,70	0,68	0,67	0,68	0,73
0,6	0,74	0,72	0,70	0,72	0,76
0,7	0,78	0,75	0,75	0,76	0,80
0,8	0,84	0,80	0,79	0,80	0,85
0,9	0,89	0,86	0,85	0,86	0,90
1	1,00	1,00	1,00	1,00	1,00

Tabell 6. *LR+ för olika korrelationskoefficienter under förutsättning att LR+ = 1 / LR–.*

Korrelations-koefficient	Sjukdomsprevalens				
	0,10	**0,30**	**0,50**	**0,70**	**0,90**
0	1	1	1	1	1
0,1	1,13	1,17	1,17	1,17	1,33
0,2	1,33	1,33	1,33	1,38	1,50
0,3	1,63	1,50	1,50	1,56	1,78
0,4	1,94	1,78	1,78	1,86	2,23
0,5	2,33	2,13	2,03	2,13	2,70
0,6	2,85	2,57	2,33	2,57	3,17
0,7	3,55	3	3	3,17	4
0,8	5,25	4	3,76	4	5,67
0,9	8,09	6,14	5,67	6,14	9
1,0	∞	∞	∞	∞	∞

∞ = oändlig