UGESKRIFT FOR LÆGER

Statusartikel

Kritisk læsning af artikler om diagnostiske studier

Mathias Weis Damkjær[1, 2], Søren Hess[3, 4, 5], Oke Gerke[6, 7], Karsten Juhl Jørgensen[1, 2] & Jeppe B. Schroll[1, 2, 8]

1) Cochrane Denmark & Centre for Evidence-Based Medicine Odense (CEBMO), 2) Open Patient Data Explorative Network (OPEN), Odense Universitetshospital, 3) Røntgen, Skanning og Nuklearmedicin, Syddansk Universitetshospital – Esbjerg Sygehus, 4) Institut for Regional Sundhedsforskning, Syddansk Universitet, 5) IRIS – Imaging Research Initiative Southwest, Esbjerg, 6) Nuklearmedicinsk Afdeling, Odense Universitetshospital, 7) Forskningsenhed for Klinisk Fysiologi og Nuklearmedicin, Klinisk Institut, Syddansk Universitet, 8) Gynækologisk Obstetrisk Afdeling, Odense Universitetshospital

Ugeskr Læger 2024;186:V06230404. doi:10.61409/V06230404

Peter er 32 år og henvender sig i almen praksis med akutte, udstrålende rygsmerter. Han er øm paravertebralt og har smerter ved strakt benløft-test, men han har ingen fokale neurologiske udfald. Arbejdsdiagnosen er akut lumbago med iskias, men vi er i tvivl om, hvorvidt Peter skal have foretaget en MR-skanning på mistanke om en lumbal diskusprolaps, da strakt benløft-test er positiv. Vi har i øvrigt ikke mistanke om alvorlig sygdom, så billeddiagnostik udlades [1], og Peters rygsmerter forsvinder i løbet af 14 dage.

HOVEDBUDSKABER

- Tolkning af test udgør en stor del af lægearbejdet og kan være vanskelig.
- Tests akkuratesse (ikkepatientrelevante endemål) estimeres i diagnostiske studier.
- Diagnostiske studier varierer meget i kvalitet, og vurdering af bias, statistisk præcision og den kliniske anvendelighed er derfor vigtigt.

MEDICINSKE TEST

Medicinske test er med til at støtte kliniske beslutninger og bruges hovedsageligt til diagnostik, men også til f.eks. stadieinddeling, monitorering og risikobestemmelse. Test kan være alt fra en klinisk undersøgelse jf. casen til en avanceret skanning. Reproducerbarheden varierer dermed, f.eks. er kliniske undersøgelser observatørafhængige, mens blodprøver kan påvirkes af måleusikkerheder. Akkuratesse beskriver testens evne til at detektere en tilstand (f.eks. sygdom), når den er til stede, og til at udelukke en tilstand, når den ikke er til stede. Akkuratessen varierer ligeledes og derfor også forskellen i risiko for falsk positive og falsk negative prøvesvar (**Tabel 1**).

TABEL 1 Oversigt over definitioner.

Term	Betydning
Akkuratesse	Testens evne til at detektere en tilstand f.eks. sygdom, når den er til stede, og til at detektere fraværet af en tilstand, når den ikke er til stede
Bias	Systematiske fejl, der gør at det observerede resultat, her sensitivitet og specificitet, afviger fra »sandheden« Afvigelsen skyldes altså ikke tilfældigheder: statistisk usikkerhed
Sensitivitet	Statistisk mål for akkuratessen af en test til at detektere syge individer i en given population
Specificitet	Statistisk mål for akkuratessen af test til at detektere raske individer i en given population
Falsk positiv	Indekstesten kategoriserer en person, som faktisk er rask, som værende syg
Falsk negativ	Indekstesten kategoriserer en person, som faktisk er syg, som værende rask
Indekstest	Den test, som vi ønsker at undersøge
Overdiagnostik	Påvisning af risikofaktorer eller diagnoser, der aldrig ville være opdaget uden testen, og som aldrig ville føre til symptomer eller død Medfører at folk unødvendigt bliver til patienter og kan føre til overbehandling med medfølgende skadevirkninger
Positiv prædikativ værdi	Sandsynligheden for, at en person faktisk er syg givet en positiv test
Negativ prædikativ værdi	Sandsynligheden for at en person faktisk er rask givet en negativ test
Referencestandard	Guldstandarden for at kategorisere syg og rask.
ROC-kurve	Receiver operating characteristic-kurve, der afbilder en numerisk tests sand positiv og negativ rate til forskellige tærskelværdier

a) Jf. konfidensintervallerne.

Derfor kan tolkning af test være vanskeligt, og prøvesvar kan potentielt lede diagnostikken og dermed valget af behandling på vildspor [2]. De kliniske oplysninger og testsvar er grundlaget for det lægefaglige skøn, hvor viden om testens akkuratesse og sygdomshyppighed inddrages.

DIAGNOSTISK AKKURATESSE

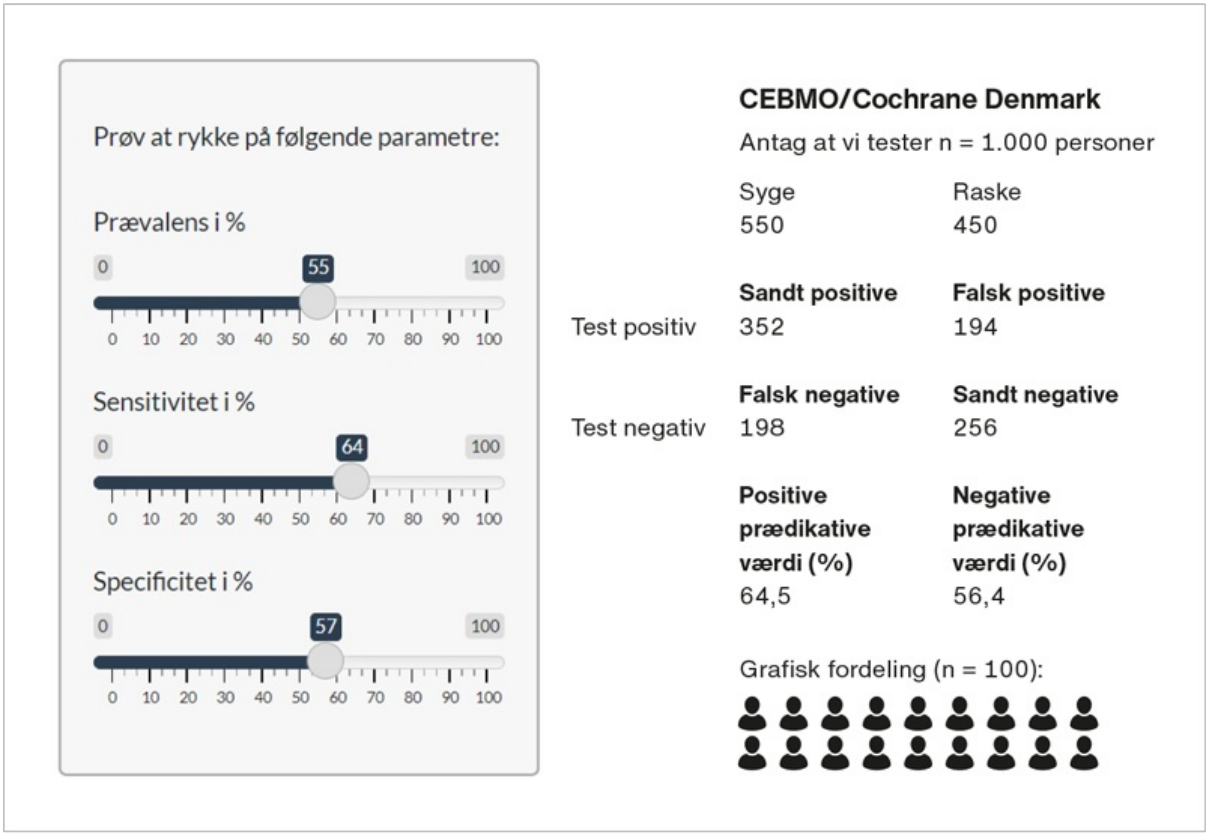
Akkuratessen kan beskrives med begreberne »sensitivitet« og »specificitet«, men der findes også andre statistiske mål [3, 4]. Sensitivitet defineres som andelen af syge, som har en sand positiv test. Specificitet defineres som andelen af raske, som har en sand negativ test (Tabel 1). Sensitivitet og specificitet estimeres i

diagnostiske studier, hvor man undersøger, hvor godt en test diskriminerer mellem syge og raske defineret ud fra en referencestandard (guldstandard). »Syg« og »rask« klassificeres med en referencestandard, som typisk vil være den test, der er bedst til bestemmelse af, om en patient har en given sygdom eller tilstand.

En tommelfingerregel angiver, at en test med høj sensitivitet er god til udelukkelse af sygdomme (få falsk negative), og en test med høj specificitet er god til bekræftelse af sygdomme (få falsk positive). Bemærk dog, at hverken sensitivitet eller specificitet beskriver forholdet mellem andelen af syge og raske. Forestil dig en veltrænet trøffelgris med en uhyre god lugtesans. Grisen vil have større sandsynlighed for at finde en trøffel i en skov i Piemonte end i kælderen på Rigshospitalet, selvom lugtesansen er den samme begge steder. Sensitivitet og specificitet knytter sig altså til testen (grisen) og sandsynligheden for at identificere sygdom afhænger, ud over testens akkuratesse, af sygdomshyppighed (flere trøfler i Piemonte). Dette er en væsentlig årsag til, at en diagnostisk test, som måske er velegnet i en selekteret population i klinikken, ikke nødvendigvis er en god screeningstest blandt den almene befolkning.

Positive og negative prædikative værdier (PPV og NPV) bruges til at estimere patientens sandsynlighed for en sygdom eller tilstand, givet et positivt eller negativt testresultat [4]. Vi har udviklet en webside [5], hvor du selv kan teste, hvad prævalensen betyder for de prædiktive værdier afhængigt af prætest (baseline) sandsynligheden (**Figur 1**). Man har dog empirisk vist, at sensitivitet og specificitet også kan påvirkes af prævalensen, muligvis indirekte gennem sværhedsgraden af sygdommen [7]. Testens akkuratesse kan derfor variere mellem forskellige populationer, f.eks. bedre i svære sygdomsstadier (på hospitalet) end i milde sygdomsstadier (almen praksis) [8].

FIGUR 1 R Shiny-applikation til illustration af statistiske mål for diagnostisk nøjagtighed [5]. Præindtastede tal er baseret på [6].



RECEIVER OPERATING CHARACTERISTIC-KURVER

Nogle test resulterer i et tal, f.eks. blodprøven C-reaktivt protein (CRP). Spørgsmålet er så, hvor sikkert CRP forudsiger bakteriel infektion, hvis CRP f.eks. er 10 mg/l eller 120 mg/l?

De enkelte tærskelværdier kan afbildes grafisk på en receiver operating characteristic (ROC)-kurve [9] med sensitivitet (sand positiv rate (y-aksen)) over 1-specificitet (falsk positiv rate (x-aksen)). Arealet under kurven (AUC) er et mål for testens akkuratesse. Er AUC 0,5, så er testens akkuratesse ikke bedre end at slå plat eller krone, mens 1 angiver den hypotetiske, perfekte test. Tærskelværdien, som optimerer forholdet mellem sensitivitet og specificitet, kan bestemmes fra ROC-kurven [4, 9].

Sensitiviteten og specificiteten påvirkes af den valgte tærskelværdi [10]. Skal flest bakterielle infektioner opdages, kan CRP > 20 mg/l vælges som tærskelværdi (høj sensitivitet). Det resulterer dog i en lavere specificitet, da mange patienter med en viral infektion vil have en CRP-værdi over 20 mg/l. Valg af tærskelværdi afhænger derfor af den specifikke sygdoms/tilstands alvorlighed, konsekvenserne ved at overse den og bivirkningerne ved den medførende behandling.

DESIGNVARIANTER

Der er stor variation i valg af studiedesign i diagnostiske studier. Det hyppigste design for diagnostiske studier er et prospektivt kohortedesign med konsekutivt inkluderede patienter, der alle får foretaget både indekstesten og referencestandarden. I diagnostiske studier undersøger man ikke, om en patient har gavn af at få foretaget en given test, da endemålet er vurdering af testens akkuratesse og ikke patientrelevante endemål som f.eks. dødelighed. Der findes dog også randomiserede diagnostiske studier, hvor patienterne bliver allokeret til to forskellige indekstest, der holdes op mod referencestandarden [11], samt diagnostiske case-kontrol-studier [12]. Randomiserede »test-treatment«-studier [13] giver mulighed for at undersøge de afledte konsekvenser af et givent testsvar og dermed undersøgelse af patientrelevante endemål. Øget testning kan også lede til fund eller diagnoser, som ikke ville have ført til symptomer i borgerens levetid, såkaldt overdiagnostik, og omfanget af dette kan også bestemmes i randomiserede studier [14, 15].

Hvis der er der risiko for overdiagnostik, bør tabellerne udvides til en 2 × 3-tabel, hvor de overdiagnosticerede tilfælde også indgår [15].

RELEVANT STUDIE

Vi vil nu gerne bestemme Peters risiko for at have en diskusprolaps efter strakt benløft-test, og først må vi vide noget mere om testen ved at besvare spørgsmålet:

»Hvad er sensitiviteten og specificiteten af strakt benløft-test til diagnosticering af lumbal diskusprolaps hos patienter med akutte rygsmerter i almen praksis, når testen holdes op imod en MR-skanning?«

To studier er blevet fundet som eksempler: *Majlesi et al* [16] og *Vroomen et al* [6], og fra det sidste studie har vi beregnet sensitiviteten og specificiteten for strakt benløft-test til at være henholdsvis 64% (95% konfidens-interval (KI): 56-71%) og 57% (95% KI: 47-66%).

I Figur 1 og **Figur 2** har vi gentaget beregningerne ved test af 1.000 fiktive personer med antagelse om en sygdomshyppighed på hhv. 10% [17] og 55% [6].

FIGUR 2 Fiktive beregninger for 1.000 testede personer, hvis sygdomshyppigheden er 10%.

Referencestandard: »den sande sygdomsstatus«
(i casen om MR-skanning)

Indeks-testen (i casen om strakt benløft-test)		**Positiv MR-skanning**	**Negativ MR-skanning**	
	Positiv strakt benløft-test	64 (a = sand positiv)	387 (b = falsk positiv)	**PPV** $\frac{a}{(a+b)} = 14\%$ 95% KI: 11-18%
	Negativ strakt benløft-test	36 (c = falsk negativ)	513 (d = sand negativ)	**NPV** $\frac{d}{(c+d)} = 93\%$ 95% KI: 91-95%
		Sensitivitet $\frac{a}{(a+c)} = 64\%$ 95% KI: 54-73%	**Specificitet** $\frac{d}{(d+b)} = 57\%$ 95% KI: 54-60%	

KI = incidensinterval; NPV = negativ prædikativ værdi; PPV = positiv prædikativ værdi.

Rettelse: Grundet satsfejl er denne figur rettet 9. januar 2024.

Vi ser her, at risikoen for diskusprolaps kun er 14% ved en positiv test ved antagelse om en sygdomshyppighed på 10%. Testen er altså ikke specielt god til at støtte vores kliniske beslutning hos patienter med lille risiko for diskusprolaps.

KRITISK VURDERING

Syv kritiske spørgsmål ved læsning af diagnostiske studier er angivet i **Tabel 2.** Punkterne omhandler manglende intern validitet (f.eks. risiko for bias) eller ekstern validitet (klinisk anvendelighed). Konsekvensen kan blive enten en over- eller underestimering af testenes nøjagtighed i en given patientgruppe.

TABEL 2 Syv kritiske spørgsmål, der kan stilles ved læsning af diagnostiske studier. For en mere fyldestgørende tjekliste anbefales brugen af STARD-tjeklisten [18] og QUADAS-2 [19].

No.	Spørgsmål
1	Er patienterne konsekutivt, prospektivt rekrutteret?
2	Har patienterne fået foretaget øvrige test, der kunne være årsag til henvisning til indekstesten og referencestandarden?
3	Er indekstesten udført korrekt blindet for referencestandarden?
4	Er referencestandarden fyldestgørende, og er den fortolket blindet for indekstesten?
5	Har alle patienter fået foretaget referencestandarden?
6	Hvor stor er den statistiske præcision?[a]
7	Er din patient og studiets patienter sammenlignelige, if. studiets inklusions- og eksklusionskriterier?

a) Jf. konfidensintervallerne.

Studiet af *Majlesi et al* [16] er et case-kontrol-studie, hvor strakt benløft-testen udføres hos to grupper af patienter, som forfatterne selv har udvalgt på baggrund af MR-svaret (referencestandarden). Designet bør generelt undgås, da det er sårbart for selektionsbias, da en højrisikopopulation typisk sammenlignes med helt raske kontrolpersoner [12]. Man risikerer herved at overestimere akkuratessen af strakt benløft-testen i forhold til en generel population [20, 21]. I studiet af *Vroomen et al* [6] er patienterne konsekutivt, prospektivt rekrutteret fra almen praksis (spørgsmål 1, Tabel 2). Patienterne henvises, hvis de har rygsmerter med udstrålende smerter til benet. Hermed undgås, at patienter udvælges pga. tidligere test (spørgsmål 2, Tabel 2). Sygdomshyppigheden af diskusprolapser i studiet er dog meget høj (ca. 55%) sammenlignet med casens (Peters) population (ca. 10%) [17].

I begge studier [6, 16] kender undersøgeren ikke resultatet af referencestandarden (MR-skanningerne). »Blindingen« er vigtig, da tolkningen af strakt benløft-testen kan påvirkes, hvis lægen har kendskab til MR-svaret, og det vil kunne give anledning til bias (spørgsmål 3, Tabel 2) [20].

MR-skanning er valgt som referencestandard i begge studier [6, 16]. Definitionen af »syg« og »rask« skal være oplyst i artiklen. I de to studier er definitionen uklar. I *Majlesi et al* [16] angives ingen definition, og i *Vroomen et al* [6] beskrives, hvilken grad af rodpåvirkning der vurderes som værende signifikant. Da nogle asymptomatiske patienter har radiologiske fund, som er forenelige med diskusprolaps, er MR-skanning alene ikke den »perfekte« referencestandard [22]. Den optimale referencestandard kunne muligvis være diskusprolaps påvist under rygkirurgisk intervention og med umiddelbar symptomlindring efterfølgende, men denne referencestandard er kun tilgængelig i en højselekteret kirurgisk population [23, 24].

I studierne [6, 16] er radiologen, der beskriver referencestandarden, blindet for kliniske oplysninger. Har radiologen kendskab til, at patienten har positiv strakt benløft-test er det muligt, at der vil være en tilbøjelighed til, at gråzonetilfælde fortolkes som syge frem for raske (dvs. større risiko for overdiagnostik). Dette vil også kunne give anledning til bias (spørgsmål 4, Tabel 2). I en klinisk virkelighed sammenholder radiologen dog altid billeddiagnostikken med kliniske oplysninger, hvilket øger overførbarheden af studiets resultater til klinikken.

Alle patienter, der har fået foretaget indekstest, bør også testes med referencestandarden (spørgsmål 5, Tabel 2). Dette kan dog være vanskeligt, særligt inden for kræftudredningsstudier, fordi referencestandarden ofte er invasiv. Nogle gange inkluderes patienterne derfor retrospektivt ud fra histologisvaret (referencestandarden), og dermed ekskluderes alle patienter, der har fået foretaget indekstesten uden referencestandarden. Det kunne være patienter, hvor mistanken ikke var stor nok til at retfærdiggøre yderligere udredning. I studier, hvor patienter rekrutteres prospektivt, vil man i visse tilfælde ikke foretage referencestandarden pga. risiko for patientskade (f.eks. grovnålsbiopsi). Dermed bliver indekstest fra disse patienter ikke verificeret mod referencestandarden, og de ekskluderes ofte fra studiet. Konsekvensen i begge scenarier bliver en mulighed for bias, da kun de svære sygdomstilfælde indgår. »Differentiel verifikation« er en måde at håndtere denne bias på ved at have en anden referencestandard for patienter med negative indekstest. Det kunne være klinisk followup, som dog er mindre præcis end f.eks. grovnålsbiopsi (referencestandarden). Dette vil medføre en ringere referencestandard i den testnegative gruppe med mulig misklassifikation af syg og rask og derved risiko for bias [25]. Ligesom i andre studier er der en statistisk usikkerhed på vores estimat af testenes akkuratesse (spørgsmål 6, Tabel 2), og det skal altid vurderes, hvordan studierne passer på den konkrete patient – ekstern validitet (spørgsmål 7, Tabel 2).

RAPPORTERINGSRETNINGSLINJER OG CRITICAL APPRAISAL TOOLS

Standards for Reporting Diagnostic Accuracy (STARD) [18], der er en rapporteringsguide, og QUADAS-2 [19], der bliver brugt til at vurdere bias og den kliniske anvendelighed af diagnostiske studier, kan benyttes til at vurdere diagnostiske studier kritisk. Redskaberne forsøger at fange elementer i studiedesign, som kan give anledning til manglende intern og ekstern validitet. Desuden har Cochrane udgivet en håndbog om review af diagnostiske studier [26].

Korrespondance *Mathias Weis Damkjær.* E-mail: mwdamkjaer@health.sdu.dk

Antaget 1. november 2023

Publiceret på ugeskriftet.dk 8. januar 2024

Interessekonflikter ingen. Forfatternes ICMJE-formularer er tilgængelige sammen med artiklen på ugeskriftet.dk

Referencer findes i artiklen publiceret på ugeskriftet.dk

Artikelreference Ugeskr Læger 2024;186:V06230404

doi:10.61409/V06230404

SUMMARY

Critical reading of scientific arcticles on diagnostic studies

Mathias Weis Damkjær, Søren Hess, Oke Gerke, Karsten Juhl Jørgensen & Jeppe B. Schroll

Ugeskr Læger 2024;186:V06230404

The issue of this review is diagnostic accuracy studies by means of which it can be determined how precisely a test can identify patients with or without a given target condition. Diagnostic accuracy studies vary substantially in design but usually report outcome measures such as sensitivity and specificity. Diagnostic accuracy studies can be critically appraised by using QUADAS-2. Some study characteristics to consider is the avoidance of a case-control study design, patient selection, and the reference standard. Also, the statistical imprecision and the applicability of a study to a general population are essential factors to consider.

REFERENCER

1. Dragsbæk L, Jensen TS, Arnbak B et al. Den kliniske relevans af MR-skanning af lænden. Ugeskr Læger. 2023;185;V03220158.
2. Graff S, Oppfeldt AM, Gotfredsen M, Christensen B. Diagnostisk bias. Ugeskr Læger. 2022;184:V06210530.
3. McGee S. Simplifying likelihood ratios. J Gen Intern Med. 2002;17(8):647-650.
4. Hróbjartsson A, Lundh A, red. Evidensbaseret medicin og klinisk forskningsmetode. Munksgaard, 2022.
5. https://sensi.shinyapps.io/sensitivity (20. nov 2023).
6. Vroomen PCAJ, Krom MCTFMD, Wilmink JT et al. Diagnostic value of history and physical examination in patients suspected of lumbosacral nerve root compression. J Neurol Neurosurg Psychiatry 2002;72(5):630.
7. Leeflang MMG, Rutjes AWS, Reitsma JB et al. Variation of a test's sensitivity and specificity with disease prevalence. CMAJ. 2013;185(11):E537-44.
8. Willis BH. Spectrum bias – why clinicians need to be cautious when applying diagnostic test studies. Fam Pract. 2008;25(5):390-6.
9. Gerke O, Zapf A. Convergence behavior of optimal cut-off points derived from receiver operating characteristics curve analysis: a simulation study. Mathematics. 2022;10(22):4206.
10. Schouten HJ, Geersing GJ, Koek HL et al. Diagnostic accuracy of conventional or age adjusted D-dimer cut-off values in older patients with suspected venous thromboembolism: systematic review and meta-analysis. BMJ. 2013;346:f2492.
11. Yang B, Mustafa RA, Bossuyt PM et al. GRADE Guidance: 31. Assessing the certainty across a body of evidence for comparative test accuracy. J Clin Epidemiol. 2021;136:146-56.
12. Rutjes AW, Reitsma JB, Vandenbroucke JP et al. Case–control and two-gate designs in diagnostic accuracy studies. Clin Chem. 2005;51(8):1335-41.
13. Hot A, Bossuyt PM, Gerke O et al. Randomized test-treatment studies with an outlook on adaptive designs. BMC Med Res Methodol. 2021;21(1):110.
14. Brodersen J, Schwartz LM, Heneghan C et al. Overdiagnosis: what it is and what it isn't. BMJ Evid Based Med. 2018;23(1):1-3.
15. Jønsson ABR, Brodersen JB. Snart er vi alle patienter. Samfundslitteratur, 2022.
16. Majlesi J, Togay H, Ünalan H et al. The sensitivity and specificity of the Slump and the Straight Leg Raising tests in patients with lumbar disc herniation. J Clin Rheumatol. 2008;14(2):87-91.
17. Kold S, Jensen AN, Christensen B. Rygsmerter. https://www.sundhed.dk/sundhedsfaglig/laegehaandbogen/fysmed-og-rehab/symptomer-og-tegn/rygsmerter/ (16. jun 2023).
18. Cohen JF, Korevaar DA, Altman DG et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open. 2016;6(11):e012799.
19. Whiting PF, Rutjes AWS, Westwood ME et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36.
20. Rutjes AWS, Reitsma JB, Nisio MD et al. Evidence of bias and variation in diagnostic accuracy studies. CMAJ.&; 2006;174(4):469-76.
21. Whiting P, Rutjes AWS, Reitsma JB et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med. 2004;140(3):189-202.
22. Jarvik JG, Hollingworth W, Heagerty PJ et al. Three-year incidence of low back pain in an initially asymptomatic cohort: clinical and imaging risk factors. Spine (Phila Pa 1976). 2005;30(13):1541-8.
23. Kim JH, van Rijn RM, van Tulder MW et al. Diagnostic accuracy of diagnostic imaging for lumbar disc herniation in adults with low back pain or sciatica is unknown; a systematic review. Chiropr Man Therap 2018;26:37.

24. Van der Windt D, Simons E, Riphagen II et al. Physical examination for lumbar radiculopathy due to disc herniation in patients with low-back pain. Cochrane Database Syst Rev 2010;2:CD007431.
25. Schmidt RL, Factor RE. Understanding sources of bias in diagnostic accuracy studies. Arch Pathol Lab Med. 2013;137(4):558-65.
26. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. https://training.cochrane.org/handbook-diagnostic-test-accuracy (16. jun 2023).