UGESKRIFT FOR LÆGER

Statusartikel

Fælles effektmål og måleinstrumenter fører til mere betydningsfuld klinisk forskning

Dorthe B. Berthelsen[1, 2, 3], Linnea Rishøj Thorlacius[1, 4, 5, 6], Bente Villumsen[4, 7], Gregor B.E. Jemec[4, 6] & Robin Christensen[1, 2, 8]

1) Parker Instituttet, Københavns Universitetshospital – Bispebjerg og Frederiksberg Hospital, 2) Forskningsenheden for Reumatologi, Klinisk Institut, Syddansk Universitet, 3) Guldborgsund Genoptræning, Guldborgsund Kommune, 4) Dermatologisk Afdeling, Sjællands Universitetshospital, Roskilde, 5) Dermatologisk Afdeling, Københavns Universitetshospital – Bispebjerg Hospital, 6) C3 – The CHORD COUSIN Collaboration, New York, 7) Patientforeningen HS Danmark, 8) Cochrane Danmark & Center for Evidensbaseret Medicin Odense (CEBMO), Klinisk Institut, Syddansk Universitet

Ugeskr Læger 2024;186:V12230799. doi: 10.61409/V12230799

HOVEDBUDSKABER

Når man designer et klinisk forskningsprojekt, bør man undersøge, om der er udviklet et egnet core outcome set (COS).

Projektet bør som minimum inkludere de effektmål og instrumenter, som COS'et beskriver.

Ved at følge COS'et muliggøres evidenssyntese, og projektet bidrager til evidensbaseret medicin.

I klinisk forskning bør man vælge effektmål, der er relevante for både patienter og klinikere [1]. I Ugeskrift for Læger beskrev vi tidligere, hvordan et effektmål, som ikke er konsensusbaseret og integreret i den øvrige litteratur, er en potentiel trussel for troværdigheden i klinisk forskning, og hvordan et såkaldt core outcome set (COS) – en minimumsliste af konsensusbaserede effektmål – sikrer ensrettede valg af effektmål [2]. Første skridt mod et COS er at blive enige om, hvad – dvs. hvilke domæner – der skal måles i et klinisk forskningsprojekt. I vores første artikel beskrev vi den globale konsensusproces, der ligger bag beslutningen om et core domain set (CDS).

Næste skridt er at opnå enighed om et core measurement set (CMS), der beskriver, hvordan – dvs. med hvilke instrumenter – man skal måle effekten på de konsensusbaserede domæner. Med »instrumenter« mener vi værktøjer, der måler kvalitet eller mængde/omfang af en variabel. Det kan f.eks. være et enkelt spørgsmål, et spørgeskema, en score opnået gennem fysisk undersøgelse, en laboratoriemåling eller et resultat vurderet ud fra et billede [3].

Til sit kliniske forskningsprojekt bør man således vælge effektmål, der er konsensusbaserede, hvor der findes et komplet COS bestående både af et »hvad«, dvs. et CDS, og et »hvordan«, dvs. et CMS. Vi beskriver her den globale konsensusproces for udvælgelse af, hvordan man måler, hvad der er vigtigt.

Konceptuelle overvejelser og identifikation af måleinstrumenter

Flere guidelines [4-6] og organisationer som Core Outcome Measures in Effectiveness Trials (COMET) og COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) [7-9] har beskrevet

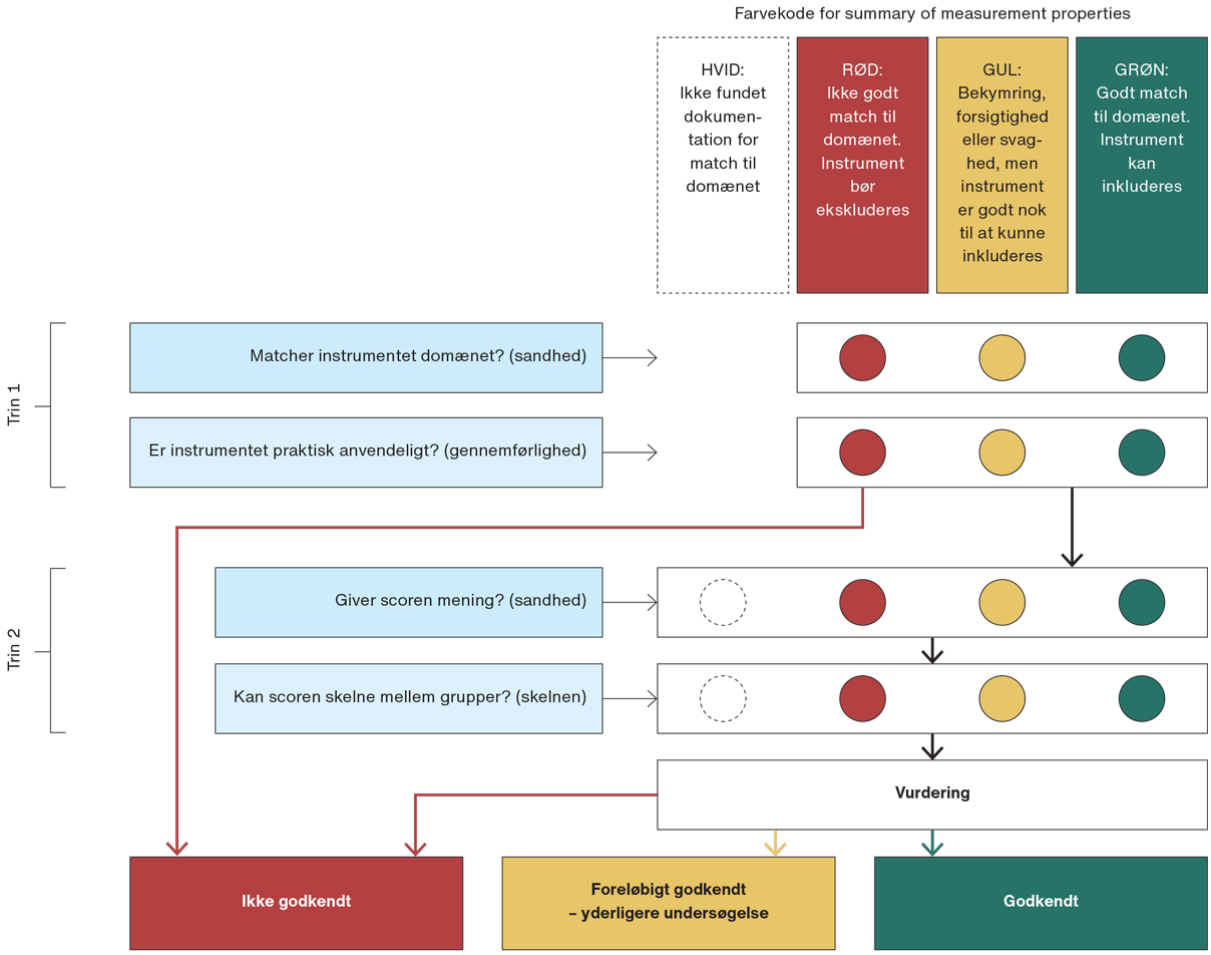
generelle metoder til at opnå konsensus om COS'er, og hvordan kvaliteten af disse vurderes. Den globale indsats på området startede inden for reumatologien i det, der i dag kendes som Outcome Measures in Rheumatology (OMERACT) [3, 10, 11]. Først bør hvert enkelt domæne i et CDS have en klar definition med målrettede og detaljerede beskrivelser af målgruppe og specifikke komponenter, der skal sikre, at et givet instrument måler det, der er intentionen [9, 12, 13]. Herefter søger man systematisk efter litteratur i flere databaser for at identificere de instrumenter, der anvendes inden for sygdomsområdet. Søgningen er en omfattende proces, der med fordel kan bygge på COSMINs vidtrækkende søgefiltre af begreber for måleegenskaber, som kombinerer begreber inden for det sygdomsområde, man undersøger [14]. Efterfølgende vurderes instrumenternes måleegenskaber [9, 15, 16].

Vurdering af instrumenters kvalitet, måleegenskaber og gennemførlighed

Der skal opnås enighed om, at et givet instrument egner sig til at måle det ønskede domæne. Konsensusprocessen bygger på evidens for, om et instrument er egnet til brug i et klinisk forsøg inden for en given sygdomsgruppe eller område. Til at indhente, vurdere og syntetisere evidensen kan man benytte guidelines og tjeklister fra COSMIN [17], der følger samme terminologi som Cochranes [18, 19] og best practice [20] (**Figur 1**). Evidensen fremlægges for at dokumentere instrumentets gyldighed ud fra dets egenskaber, f.eks. validitet og pålidelighed. Processen foregår i to trin og kan beskrives ud fra tre begreber: sandhed, gennemførlighed og skelnen [3, 13] (**Figur 2**).

FIGUR 1 Elementer, som indhentning og syntese af evidensen bygger på.

FIGUR 2 Udvælgelsesproces for hvert enkelt instrument, hvor fire spørgsmål besvares med en farvekode for at vurdere instrumentets sandhed, gennemførlighed og skelnen [13].



Trin 1: Domænematch og brugervenlighed

Først ser man på, hvor godt et instrument passer til det, man ønsker at måle, og hvor brugervenligt det er. Man kan opstille resultaterne af domænematchet og brugervenligheden i en summary of measurement properties (SOMP)-tabel med en faciliterende farvekode [13] (Figur 2).

Matcher instrumentet domænet?

Ud fra domænets definition vurderer man først, om et givet instrument er et godt match til domænet, og om instrumentets indhold beskriver oplevelsen af det pågældende domæne i den tilsigtede målgruppe og undersøgelsessituation [21]. Man ser på instrumentets indhold og formål ved at undersøge to egenskaber [9, 13, 22, 23]: 1) indholdsvaliditet: I hvilket omfang indholdet i instrumentet afspejler det koncept, som ønskes målt; 2) overfladevaliditet: I hvilken grad ser instrumentet som helhed ud til at matche domænet?

Til vurderingen kan man anvende forskellige retningslinjer og kriterier fra f.eks. COSMIN, som arbejder for bedre udvælgelse af måleinstrumenter [24]. COSMIN har f.eks. udviklet en tjekliste for patientrapporterede måleinstrumenter, der bl.a. vurderer, om indholdet er relevant for domæne, målgruppe og kontekst, om instrumentet dækker alle ønskede begreber, og om målgruppen forstår alle spørgsmål i instrumentet på den måde, man har intentioner om [25]. Når man skal vurdere, om instrumentet er et godt match til domænet, er det vigtigt at involvere centrale interessenter, herunder patienter med lidelsen, så man sikrer, at instrumentets indhold passer til målgruppen [8, 13]. Det kan være vigtigt selv at undersøge eller se på andre undersøgelser,

som viser fordelingen af svar, tilbøjelighed til manglende svar eller såkaldte floor- og ceiling-effekter, dvs. om mange deltageres svar har tendens til »at klumpe sig sammen« i bunden eller i toppen – noget, der kan være tegn på problemer med, at indhold og svarkategorier passer til målgruppen [13].

Er instrumentet brugervenligt?

Næste skridt er at vurdere instrumentets brugervenlighed. Man ser her på, om instrumentet er praktisk at anvende ved at vurdere elementer som omkostninger, tilgængelighed på det eller de nødvendige sprog, behov for oplæring, længde, tid- og energiforbrug, administration og anden byrde, som anvendelse af instrumentet medfører for både forskeren og klinikeren og for patienten [22, 26]. Der er derfor ikke bare behov for at involvere sundhedspersoner, men også patienter med pågældende lidelse, som kan vurdere, om instrumentets format er egnet [13].

Hvis flere instrumenter matcher indholdet i et domæne, kan det være relevant at lave en spørgeskemaundersøgelse, hvor man f.eks. beder patienter i målgruppen om at vurdere og sammenligne elementer som forståelighed, tidsforbrug og svarmuligheder [27].

Konsensus: Er instrumentet både et match med domænet og brugervenligt?

Ud fra vurderingen af domænematch og brugervenlighed skal arbejdsgruppen nå til enighed om, hvilke instrumenter der fortsat skal inkluderes, og hvilke der kan ekskluderes [8, 13]. Man bør kun fokusere på instrumenter, der har en grøn eller gul vurdering i SOMP'en (Figur 2). Ofte kan man eliminere adskillige instrumenter, der ikke dækker det rette indhold til den påtænkte undersøgelsessituation eller virker for lange, dyre og/eller komplekse.

Trin 2: Gennemgang af evidens for instrumenters måleegenskaber

Hvis instrumentet er både et godt match og er brugervenligt, vurderer man instrumentets måleegenskaber. Når man gennemgår egenskaberne, fortsætter man arbejdet i SOMP'en, hvor hver enkelt måleegenskab undersøges individuelt.

Giver scoren mening?

Først ser man på, om måleskalaen giver mening, så skalaen afspejler det indhold, man oprindeligt havde beskrevet, at domænet skulle måle. Man vurderer, om skalaen giver mening, ved at undersøge tre egenskaber [9, 13, 22, 23]: 1) Gyldighed: Begrebsvaliditet beskriver, i hvilken grad instrumentets scorer stemmer overens med scorerne på sammenlignelige instrumenter, der måler på teoretisk afledte, tilsvarende domæner eller viser en forskel i score mellem grupper med samme kendte forskel. I nogle tilfælde taler man om en guldstandard, dvs. kriterievaliditet, som andre instrumenter kan vurderes op i mod; som udgangspunkt findes guldstandarder ikke i psykometrien, måske undtaget et ønske om at validere en forkortet udgave af et spørgeskema mod det originale. 2) Pålidelighed: i hvilken grad instrumentets score er fri for målefejl. Ved intrareliabilitet ser man på instrumentets evne til at give ensartede scorer, f.eks. når den samme bedømmer med passende mellemrum benytter instrumentet på de samme patienter. Ved interreliabilitet ser man f.eks. på instrumentets evne til at give ensartede scorer på tværs af forskellige bedømmere. Man ser også på instrumentets test-retest-reliabilitet, dvs. i hvilken grad instrumentet er reproducerbart – altså har evnen til at give ensartede scorer over tid i en stabil population. Endelig vurderer man, i hvor høj grad instrumentets elementer er relaterede til hinanden, dvs. er internt konsistense. 3) Ændring over tid: i hvilken grad et instrument kan detektere ændringer, som reelt opstår over tid.

Er ændring i scoren betydningsfuld?

Næste skridt er at vurdere, om ændring i instrumentets score er betydningsfuld ved sammenligning af grupper i

et klinisk forsøg. Her undersøger man fire egenskaber [9, 13, 22, 23]: 1) Velkendt grænse for en meningsfuld ændring: I hvilken grad en ændring i scoren er meningsfuld, både for den enkelte patient og ved gennemsnitsberegninger. 2) Minimal detectable change: På gruppeniveau er dette næppe klinisk betydningsfuldt, da det repræsenterer den mindste ændringsværdi, der kan antages at være ud over målefejlen for instrumentet [28]. 3) Minimal important difference: Denne egenskab kan frit oversættes til den ønskede gruppeforskel [29]; dette repræsenterer den ønskede forskel i resultater mellem to (eller flere) grupper på tidspunktet for resultatvurderingen [28]. 4) Minimal important change: Dette repræsenterer den ændring, der er observeret over en periode hos en enkelt person, som anses for at være klinisk meningsfuld, svarende til at patienten har responderet på behandling [28].

For hver måleegenskab leder man efter resultater fra mindst to undersøgelser af høj kvalitet med lav risiko for bias. Undersøgelserne skal have vist, at instrumentets egenskaber er tilfredsstillende, og ingen undersøgelser må tyde på, at instrumentets måleegenskaber ikke fungerer. I praksis kan der være »huller i evidensen«, og derfor kan der være behov for selv at udføre ekstra valideringsstudier, som så skal vurderes kritisk af andre forskere for at retfærdiggøre konklusionerne [13].

Konsensus: Hvilke måleinstrumenter skal man vælge, hvis man vil lave betydningsfuld klinisk forskning?

Det er en balancegang mellem, hvorvidt instrumentet er praktisk anvendeligt, og at det skal have tilstrækkeligt indhold til at fange hele spektret af det ønskede domæne [13]. Ofte mangler der evidens for måleegenskaberne [30]. Man kan så overveje midlertidigt at benytte et brugervenligt instrument med høj evidens for god indholdsvaliditet, indtil man får valideret instrumentet fuldt ud. Findes et sådant ikke, kan det være nødvendigt at udvikle et helt nyt instrument [9]. Hvis dette er tilfældet, gennemføres endnu en konsensusproces, efter udviklingen er afsluttet [7, 8, 13]. For at opnå konsensus er det vigtigt at sikre, at synspunkter fra alle centrale interessenter er inkluderet. Det kan derfor være en fordel at anvende en konsensusmetode med en anonym afstemning, f.eks. via digitale metoder gennem apps eller hjemmesider [7].

Diskussion

Et COS er fuldt udviklet, når det består af både et CDS – dvs. hvad skal måles? – og et CMS – dvs. hvordan skal det måles? Det kan være en lang proces at nå frem til en beslutning om et COS og derefter få det implementeret. Nogle vil mene, at et COS kan mangle fleksibilitet, kræve ekstra ressourcer eller overforsimple resultater, men fordelene ved et COS er, at de bygger på en gennemarbejdet konsensusproces, og at standardisering gør det muligt at sammenligne og kombinere data på tværs af undersøgelser og derved fremme evidensbaseret medicin.

Konklusion

Når man laver klinisk forskning, bør man undersøge, om der er udviklet et COS for netop den sygdom eller det område, man vil undersøge. Finder man et egnet COS og anvender dets effektmål og måleinstrumenter, gerne rapporteret blandt »key secondary outcomes«, når man skriver sin forskningsartikel, bidrager projektet til, at den kliniske forskning bliver mere betydningsfuld, idet resultater hermed kan sammenlignes, og evidensen syntetiseres.

Korrespondance *Dorthe B. Berthelsen*. E-mail: dorthe.bang.berthelsen@regionh.dk

Antaget 23. juli 2024

Publiceret på ugeskriftet.dk 9. september 2024

Interessekonflikter Der er anført potentielle interessekonflikter. Forfatternes ICMJE-formularer er tilgængelige sammen med artiklen på ugeskriftet.dk

Referencer findes i artiklen publiceret på ugeskriftet.dk

Artikelreference Ugeskr Læger 2024;186:V12230799

doi 10.61409/V12230799

SUMMARY

Common efficacy measures and measurement instruments lead to more meaningful clinical research

This review describes that a core outcome set (COS) represents a consensus-based minimum set of outcomes to be collected and reported in clinical trials involving a particular disease or population. A COS serves as a guideline for global consensus on which outcome domains should be collected in all clinical trials. After defining what to measure, it becomes crucial to reach consensus on how to measure it. This includes the selection of appropriate outcome measurement instruments with credible measurement properties and interpretable thresholds of meaning.

REFERENCER

1. Christensen R, Nielsen SM, Henriksen M. Typer af data og typer af effektmål i klinisk forskning. I: Hróbjartsson A, Lundh A (red.). Evidensbaseret medicin og klinisk forskningsmetode. 1. udgave. Munksgaard, 2022:89-102.
2. Thorlacius LR, Villumsen B, Christensen R et al. Troværdig klinisk forskning kræver konsensus om effektmål. Ugeskr Læger. 2023;185:V09220579.
3. Boers M, Kirwan JR, Wells G et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. J Clin Epidemiol. 2014;67(7):745-53.
4. Kirkham JJ, Davis K, Altman DG et al. Core outcome set-STAndards for development: the COS-STAD recommendations. PLoS Med. 2017;14(11):e1002447. https://doi.org/10.1371/journal.pmed.1002447
5. Kirkham JJ, Gorst S, Altman DG et al. Core outcome set-STAndardised protocol items: the COS-STAP statement. Trials. 2019;20(1):116. https://doi.org/10.1186/s13063-019-3230-x
6. Kirkham JJ, Gorst S, Altman DG et al. Core outcome set-STAndards for reporting: the COS-STAR statement. PLoS Med. 2016;13(10):e1002148. https://doi.org/10.1371/journal.pmed.1002148
7. Williamson PR, Altman DG, Blazeby JM et al. Developing core outcome sets for clinical trials: issues to consider. Trials. 2012;13:132. https://doi.org/10.1186/1745-6215-13-132
8. Williamson PR, Altman DG, Bagley H et al. The COMET Handbook: version 1.0. Trials. 2017;18(Suppl 3):280. https://doi.org/10.1186/s13063-017-1978-4
9. Prinsen CAC, Vohra S, Rose MR et al. Guideline for selecting outcome measurement instruments for outcomes included in a core outcome set, 2016. https://www.cosmin.nl/wp-content/uploads/COSMIN-guideline-selecting-outcome-measurement-COS.pdf (2. aug 2024).
10. Tugwell P, Boers M, Brooks P et al. OMERACT: an international initiative to improve outcome measurement in rheumatology. Trials. 2007;8:38.
11. Boers M, Beaton DE, Shea BJ et al. OMERACT Filter 2.1: Elaboration of the Conceptual Framework for Outcome Measurement in Health Intervention Studies. J Rheumatol. 2019;46(8):1021-1027. https://doi.org/10.3899/jrheum.181096
12. Beaton D, Maxwell LJ, Grosskleg S et al. Chapter 4: Developing core domain sets. I: The OMERACT Handbook, Version 2.1, 2021. https://omeract.org/handbook/ (2. aug 2024).
13. Beaton D, Maxwell LJ, Grosskleg S et al (red.). Chapter 5: Instrument selection for core outcome sets. I: The OMERACT Handbook, Version 2.1, 2021. https://omeract.org/handbook/ (2. aug 2024).
14. COSMIN. Search filters, 2024. https://www.cosmin.nl/tools/pubmed-search-filters/ (8. maj 2024).

15. Terwee CB, Jansma EP, Riphagen II, de Vet HCW. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res. 2009;18(8):1115-23. https://doi.org/10.1007/s11136-009-9528-5
16. Prinsen CAC, Vohra S, Rose MR et al. How to select outcome measurement instruments for outcomes included in a "core outcome set" – a practical guideline. Trials. 2016;17:449.
17. Cosm Checkl Assess Study Qual. https://www.cosmin.nl/tools/checklists-assessing-methodological-study-qualities/ (5. dec 2023).
18. Ghogomu EAT, Maxwell LJ, Buchbinder R et al. Updated method guidelines for cochrane musculoskeletal group systematic reviews and metaanalyses. J Rheumatol. 2014;41:194-205.
19. Higgins JPT, Thomas J, Chandler J et al. Cochrane Handbook for Systematic Reviews of Interventions. https://training.cochrane.org/handbook (26. jul 2023).
20. Slavin RE. Best evidence synthesis: an intelligent alternative to meta-analysis. J Clin Epidemiol. 1995;48(1):9-18. https://doi.org/10.1016/0895-4356(94)00097-a
21. Beaton DE, Maxwell LJ, Shea BJ et al. Instrument selection using the OMERACT Filter 2.1: the OMERACT Methodology. J Rheumatol. 2019;46(8):1028-1035. https://doi.org/10.3899/jrheum.181218
22. Reeve BB, Wyrwich KW, Wu AW et al. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. Qual Life Res. 2013;22(8):1889-905. https://doi.org/10.1007/s11136-012-0344-y
23. Bombardier C, Tugwell P. Methodological considerations in functional assessment. J Rheumatol Suppl. 1987;14 Suppl 15:6-10.
24. Mokkink LB, Prinsen CAC, Bouter LM et al. The COnsensus-based standards for the selection of health measurement instruments (COSMIN) and how to select an outcome measurement instrument. Braz J Phys Ther. 2016;20(2):105-13. https://doi.org/10.1590/bjpt-rbf.2014.0143
25. Terwee CB, Prinsen CAC, Chiarotto A et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res. 2018;27(5):1159-1170. https://doi.org/10.1007/s11136-018-1829-0
26. Auger C, Demers L, Swaine B. Making sense of pragmatic criteria for the selection of geriatric rehabilitation measurement tools. Arch Gerontol Geriatr. 2006;43(1):65-83. https://doi.org/10.1016/j.archger.2005.09.004
27. Tang K, Beaton DE, Lacaille D et al. Sensibility of five at-work productivity measures was endorsed by patients with osteoarthritis or rheumatoid arthritis. J Clin Epidemiol. 2013;66(5):546-56. https://doi.org/10.1016/j.jclinepi.2012.12.009
28. Sabah SA, Alvand A, Beard DJ, Price AJ. Minimal important changes and differences were estimated for Oxford hip and knee scores following primary and revision arthroplasty. J Clin Epidemiol. 2022;143:159-68. https://doi.org/10.1016/j.jclinepi.2021.12.016
29. Cook JA, Julious SA, Sones W et al. DELTA2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. Trials. 2018;19(1):606. https://doi.org/10.1186/s13063-018-2884-0
30. Højgaard P, Klokker L, Orbai A-M et al. A systematic review of measurement properties of patient reported outcome measures in psoriatic arthritis: A GRAPPA-OMERACT initiative. Semin Arthritis Rheum. 2018;47(5):654-665. https://doi.org/10.1016/j.semarthrit.2017.09.002

