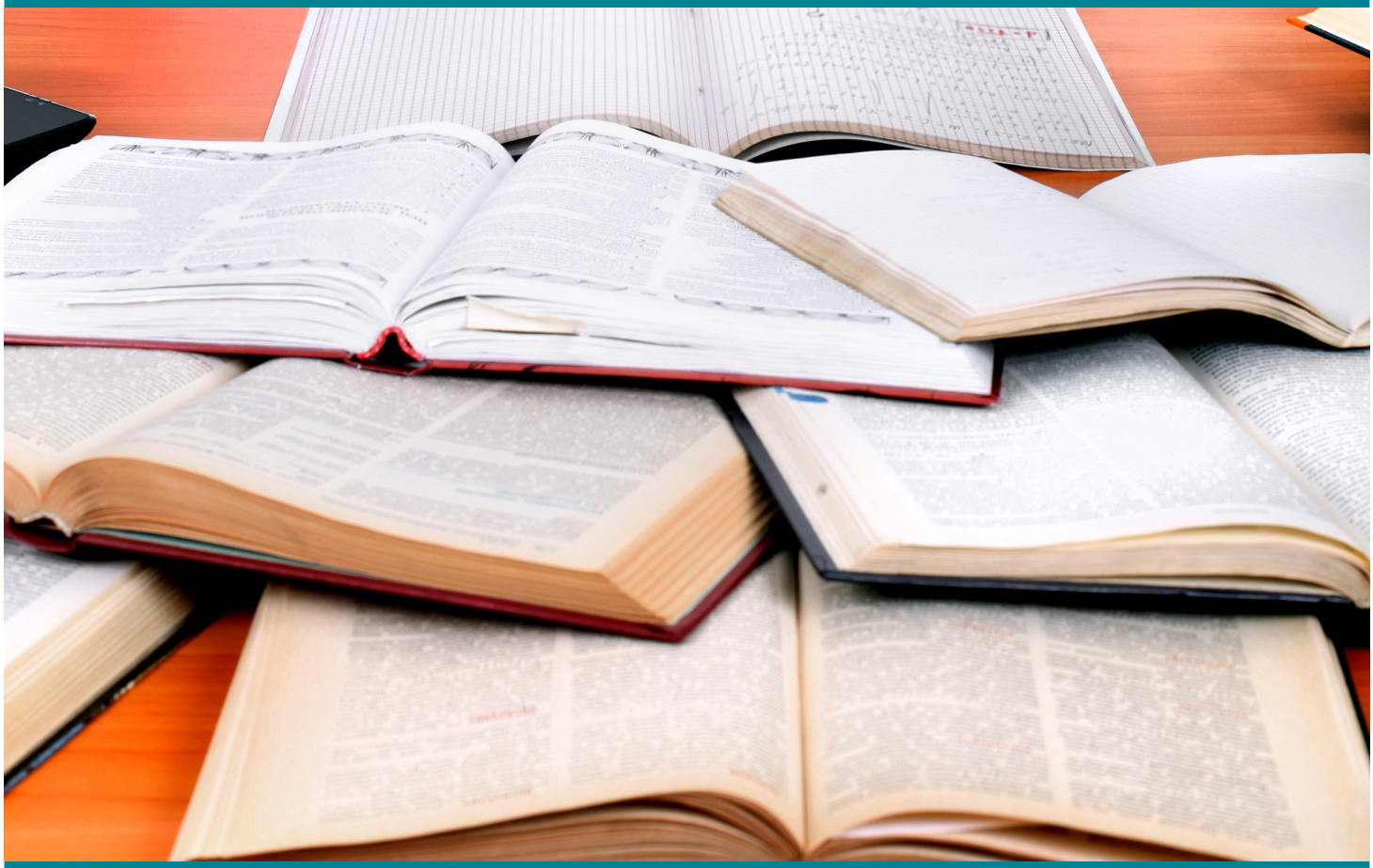
# Systematisk litteraturgennemgang

## Health Technology Assessment – Metodevejledning, vol. 1



DEFACTUM

**Systematisk litteraturgennemgang – Health Technology Assessment – Metodevejledning, vol. 1**



Emneord: Systematisk litteraturgennemgang, medicinsk teknologivurdering, MTV.

Sprog: Dansk

Udgivet af: DEFACTUM®, november 2019
Udgave: 1. udgave
Forside: Colourbox

ISBN: 978-87-93657-10-6

Denne publikation citeres således:
Valentin G, Palmhøj Nielsen C, Lou S, Groth Jensen L, Løvschall C.
Systematisk litteraturgennemgang – Health Technology Assessment – Metodevejledning, vol. 1. Aarhus: DEFACTUM, Region Midtjylland, 2019.

Kvalitetssikring: Peer-review ved Robin Christensen, leder af Musculoskeletal Statistics Unit, Parker Instituttet, Bispebjerg og Frederiksberg Hospital.



For yderligere oplysninger rettes henvendelse til:
DEFACTUM
Olof Palmes Allé 15
8200 Aarhus N

E-mail: defactum@rm.dk
Hjemmeside: www.defactum.dk

Rapporten kan downloades fra www.defactum.dk.

# Forord

Foreliggende metodevejledning for litteraturgennemgang og vurdering af evidens indeholder en beskrivelse af de fremgangsmåder og metoder, vi anvender, når vi vurderer og opsummerer kvantitativ forskning.

Systematisk opsummering af forskning kendetegnes ved en systematisk og transparent fremgangsmåde til at formulere forskningsspørgsmål, søge efter litteratur, vurdere, sammenfatte og præsentere forskningsbaseret viden. Metoderne som anvendes inden for dette felt er i kontinuerlig udvikling. I DEFACTUM, Folkesundhed og Sundhedstjenesteforskning holder vi os løbende opdateret på metodefeltet, og vi følger internationale principper for, hvordan systematiske oversigtsartikler bør udarbejdes[1]. Formålet med denne metodevejledning er på et overordnet plan at synliggøre de metoder, vi anvender. Vejledningen inkluderer ikke en detaljeret beskrivelse af de interne procedurer og arbejdsgange ved udarbejdelse af systematiske litteraturgennemgange. Målgruppen for vejledningen er primært kollegaer eller andre fagpersoner, som er involveret i eller har interesse i systematiske litteraturgennemgange. For en mere dybdegående introduktion til de metoder, vi anvender, henvises til Cochrane-håndbogen (http://training.cochrane.org/handbook) og en artikelserie om GRADE (http://www.gradeworkinggroup.org/).

[1] https://training.cochrane.org/handbook

# Indhold



**Figurer**



**Tabeller**

# 1 Forskellige typer af litteraturgennemgange

I DEFACTUM, Folkesundhed og Sundhedstjenesteforskning udarbejdes alle systematiske litteraturgennemgange på en systematisk og transparent måde. Det skal være muligt at efterprøve og kritisere vores resultater og konklusioner. Derfor anvender vi transparente og på forhånd definerede, protokollerede metoder, når vi indsamler, vurderer og sammenfatter forskningsresultater. Vi udarbejder forskellige typer af litteraturgennemgange, afhængigt af hvad opdragsgiveren ønsker, hvad produktet skal bruges til og afhængigt af tidsrammen og budgettet for arbejdet. Vi forsøger at balancere ønsket om hurtig leverance med nødvendigheden af at følge strenge, metodiske krav.

I det følgende præsenteres de forskellige typer af litteraturgennemgange, vi udarbejder.

## 1.1 Medicinske teknologivurderinger

Medicinsk Teknologivurdering (MTV) er en flervidenskabelig og tværfaglig aktivitet, som leverer input til prioriteringer og beslutningstagning i sundhedsvæsenet i relation til forebyggelse, diagnosticering, behandling og rehabilitering (SST). MTV defineres som en alsidig systematisk vurdering af forudsætningerne for og konsekvenserne af at anvende en medicinsk teknologi. I MTV-sammenhæng forstås begrebet teknologi bredt som procedurer og metoder til forebyggelse, diagnosticering, behandling, pleje og rehabilitering inkl. apparatur og lægemidler [1]. Organisering og understøttende systemer kan ligeledes ses som en medicinsk teknologi. Litteraturgrundlaget i en MTV kan være primære studier og/eller systematiske oversigtsartikler. Vurdering af effekt og sikkerhed af en given teknologi følger de samme metoder som beskrevet for systematiske litteraturgennemgange generelt.

I DEFACTUM arbejder vi med MTV på tre forskellige niveauer: Internationale, nationale og regionale MTV'er:

- **Internationale MTV'er** bliver udarbejdet i regi af EUnetHTA, som er et formaliseret samarbejde mellem ca. 60 europæiske MTV-institutioner. DEFACTUM varetager tovholderfunktionen for det tværregionale MTV-arbejde i Danmark og repræsenterer således Danmark i det internationale MTV-arbejde i EUnetHTA. Internationale MTV'er bliver udarbejdet i overensstemmelse med EUnetHTA's Core Model, som er en metodisk ramme for produktion og deling af MTV-arbejde. For yderligere information om EUnetHTA's Core Model henvises til EUnetHTA's hjemmeside: http://www.eunethta.eu/hta-core-model.

- **Nationale MTV'er** bliver almindeligvis udarbejdet efter opdrag fra de fem regioner i Danmark. DEFACTUM, Region Midtjylland planlægger og beskriver som projektudfører i tæt samarbejde med regionerne og Danske Regioner rammerne for projektarbejdet til godkendelse i sundhedsdirektørkredsen. Projektets rammer udstikkes desuden i tæt sparring med en faglig projektgruppe samt en følgegruppe bestående af relevante faglige selskaber, patientforeninger og andre relevante interessenter.
- **Regionale/lokale MTV**-produktioner kan blive igangsat efter administrativt/politisk ønske i en region. Efterfølgende iværksættes samme processer som ved nationale MTV'er.

## 1.2 Systematiske oversigter over primære studier (systematiske reviews)

Vi udarbejder systematiske litteraturgennemgange af primær forskning i de tilfælde, hvor der ikke foreligger et opdateret systematisk review af god kvalitet, eller i de tilfælde hvor der er behov for at opdatere vidensgrundlaget, da litteratursøgning gerne ligger mindst et år tilbage. Vi følger som udgangspunkt princippet om at anvende evidens af så høj kvalitet som muligt, hvilket indebærer, at vi søger efter randomiserede, kontrollerede studier (eng. Randomised Controlled Trials, kaldet RCT), når effekten af et givet tiltag i sundhedsvæsenet vurderes. Såfremt der ikke foreligger RCT'er inden for det givne område, eller såfremt problemstillingen kræver det, vil der blive søgt efter andre studiedesigns [2] som fx kontrollede før- og efter-studier, kohortestudier, case-kontrol-studier eller tværsnitsstudier. I forbindelse med fx diagnostiske review søges RCT'er og kohortestudier dog oftest ligeværdigt. Reviewets karakter er således bestemmende for den metodiske tilgang her for at benytte ressourcerne bedst muligt [2].

## 1.3 Systematiske oversigter over systematiske reviews (overview of reviews)

Et 'overview of reviews' er en term, der dækker systematisk gennemgang af systematiske reviews eller metaanalyser i modsætning til gennemgang af primære studier. Det kan være relevant at overveje at udarbejde et overview af reviews når:

- der er publiceret mange systematiske reviews, som besvarer det samme forsknings-spørgsmål.
- man ønsker at undersøge effekten af forskellige interventioner til den samme helbreds-tilstand.

- man ønsker at undersøge den samme intervention og den samme helbredstilstand, men med forskellige effektmål.
- man ønsker at undersøge den samme intervention til forskellige patientgrupper eller populationer.

Systematiske oversigter over systematiske reviews er velegnede til at give klinikere og beslutningstagere et samlet overblik over den tilgængelige evidens inden for et givet område.

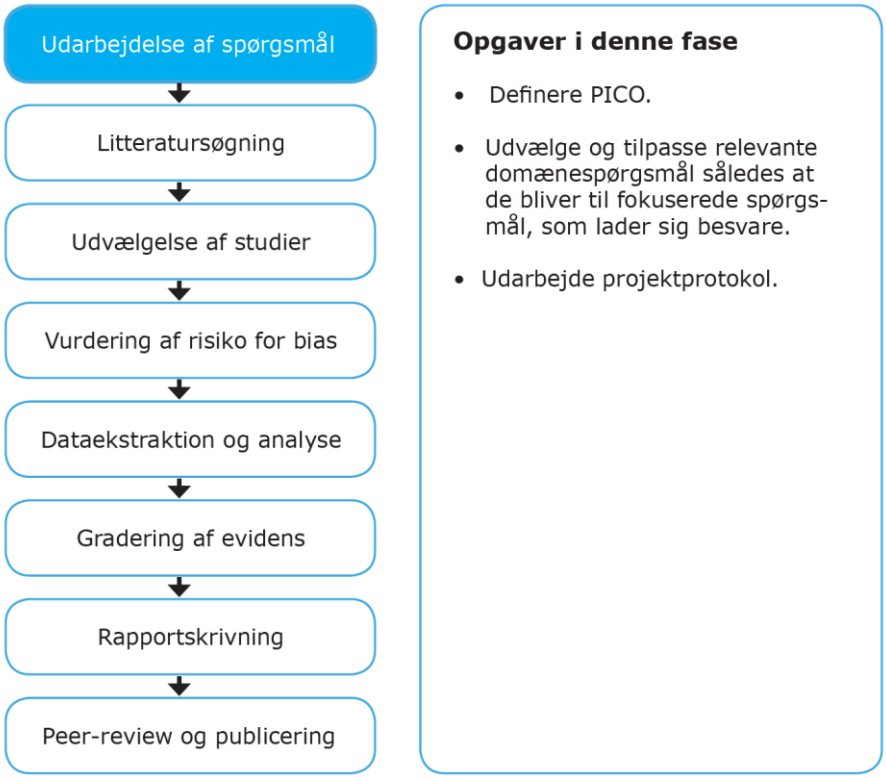
## 1.4 Omtaler af forskningsresultater

I DEFACTUM formidler vi relevante internationale forskningsorganisationers forskning via korte sammendrag. Dette gøres med henblik på at optimere udbredelsen af internationale organisationers forskning og derved mindske ressourceanvendelsen i forbindelse med udarbejdelse af MTV og systematiske litteraturgennemgange. Resultater, som er relevante, og som har nyhedsværdi i en dansk kontekst, formidles. Omtalerne af resultaterne bliver publiceret på DEFACTUMs hjemmeside i et kort sammendrag (3-4 sider). Sammenfatningerne bliver vurderet af danske fagfolk på området inden publicering. Omtale af forskningsresultater er en samlet betegnelse for omtale af forskellige publikationstyper: Omtale af en EUnetHTA-rapport (HTA Core Model), omtale af nordiske organisationers MTV-resultater, omtale af relevante systematiske reviews eller overviews (ofte Cochrane-reviews).

# 2 Overblik: Sådan udarbejder vi systematiske litteratur-studier

## 2.1 Forberedelsesfase

**Figur 1. Opgaver i forberedelsesfasen***




*Figuren er bearbejdet på baggrund af figurer i 'Håndbok for Nasjonalt kunnskapssenter for helsetjenesten desember 2015' – 'Slik oppsummerer vi forskning'.

### 2.1.1 Udvælgelse og tilpasning af fokuserede spørgsmål ved hjælp af PICO og udarbejdelse af projektprotokol

Første skridt i udarbejdelsen af en systematisk litteraturgennemgang er at få specificeret, hvad det præcist er, der skal undersøges (afgrænse emnet). Dette gøres ved at udarbejde fokuserede spørgsmål på baggrund af det overordnede forskningsspørgsmål (formål) og PICO [3]. De fokuserede spørgsmål er centrale for at kunne udarbejde en præcis søgestrategi og for at kunne foretage en systematisk udvælgelse af de identificerede referencer. Ved udarbejdelsen af MTV-rapporten tages der udgangspunkt i de prædefinerede, generiske spørgsmål fra EUnetHTA's Core Model (http://www.eunethta.eu/hta-core-model). Domænespørgsmål, som vurderes at være relevante i den givne kontekst, udvælges. Herefter tilpasses de udvalgte spørgsmål ved hjælp af det overordnede forskningsspørgsmål samt PICO-strukturen. PICO er et akronym, som står for:

- Population: Den specifikke population (patientgruppe) som teknologien henvender sig til
- Intervention: Den nye teknologi som skal vurderes
- Komparator: Sammenligningsalternativet, fx aktuel klinisk praksis
- Outcome: Effektmål

I de følgende afsnit beskrives relevante overvejelser i relation til elementerne i PICO ved vurdering af en teknologi.

### 2.1.2 Population

Den specifikke population (patientgruppe), som teknologien henvender sig til, skal defineres og beskrives. Beskrivelsen af patientgruppen (målpopulationen) omfatter præcisering af alder, køn, helbredstilstand, relevant komorbiditet samt prognose (ubehandlet og behandlet med den mest gængse behandling). Derudover beskrives incidensen og prævalensen af den givne helbredstilstand i Danmark. Hvis der forventes at være betydelige forskelle i effekt (eller omkostninger) ved behandling med den givne teknologi til særlige subgrupper, bør disse beskrives [4].

### 2.1.3 Intervention

Der skal gives en præcis og dækkende beskrivelse af den teknologi, som vurderes. Teknologien skal afgrænses og defineres ud fra sin materielle natur, sit formål og graden af udbredelse og modenhed, således at det fremgår klart, hvilket apparatur, intervention, lægemiddel eller procedure der vurderes [1]. Hvilke oplysninger der skal indgå i beskrivelsen afhænger af hvilken type af teknologi, der undersøges. Er der tale om en konkret teknologi bør beskrivelsen blandt andet omfatte, hvilken fase teknologien er i, hvilke indikationer teknologien har opnået markedsgodkendelse eller CE for, samt hvilke påståede sundhedseffekter der er ved den givne teknologi [4]. Er det i stedet en diagnostisk test, man ønsker at undersøge, bør beskrivelsen blandt andet omfatte specifikation af tærskelværdier, beskrivelse af hvem der skal udføre og fortolke testen samt beskrivelse af omgivelserne (laboratorium, klinik) [5].

### 2.1.4 Komparator (sammenligningsalternativ)

Effekten af den nye teknologi estimeres ved at vurdere, hvorvidt den givne teknologi er forbundet med en merværdi for patienten, fx i form af forbedret livskvalitet eller forlænget levetid. Merværdien eller den relative effekt af en given teknologi afhænger naturligvis af, hvilken intervention der sammenholdes med. Derfor er det essentielt, at effekten af den givne teknologi sammenholdes med et relevant alternativ. Det alternativ, der sammenholdes med, bør være det mest aktuelle alternativ i dansk kontekst. Det mest aktuelle alternativ vil være det som den nye teknologi vil erstatte (på grund af superioritet), såfremt denne teknologi tages i brug. Sammenligningsalternativet bør være gængs, klinisk praksis, da dette sammenligningsgrundlag vil være det mest relevante og informative [6]. Valget af sammenligningsalternativ(er) skal begrundes tydeligt i rapporten. Således handler det om at spørge retorisk "Behandlingen virker, men bedre end hvad?".

### 2.1.5 Outcome (effektmål)

Identifikation og udvælgelse af relevante effektmål er et vigtigt element i udarbejdelsen af et litteraturstudie. I det følgende beskrives faktorer, som er relevante at overveje, når relevante effektmål identificeres og udvælges. Dernæst følger en beskrivelse af, hvordan relevante effektmål rangordnes efter, hvorvidt de er kritiske, vigtige eller ikke vigtige for beslutningstagning. Heraf fremgår det, at der ofte er anvendt effektmål i RCT'er, der ikke har nogen værdi i en vurdering af en teknologi på trods af anerkendte forskeres store interesse i disse (3).

Kliniske effektmål har til formål at kvantificere behandlingseffekten i forhold til, hvordan patienten har det (fx livskvalitet), fungerer (fx funktion) og overlever. Denne effekt kan enten være gavnlig i form af en forbedret helbredsstatus (overlevelse, helbredelse, remission) eller skadelig (bivirkninger, indlæggelser, død). Et givet effektmål skal være klinisk relevant, validt, pålideligt og tilstrækkeligt følsomt til at kunne registrere ændringer over tid. Derudover skal effektmålet helst være bredt accepteret og anvendt af klinikere og være relevant for patienterne. En klar definition på det valgte effektmål, herunder angivelse af pålidelighed, validitet samt den statistiske og kliniske relevans, skal beskrives [7].

Kliniske effektmål kan overordnet inddeles i tre kategorier: Mortalitet, morbiditet (symptomer og bivirkninger) og helbredsrelateret livskvalitet

Ud over at få afklaret, hvilke effektmål (eng. outcome domains) der er relevante for vurderingen, skal man også gøre sig klart, hvordan disse effekter kan måles. Hvis der fx er flere

måleredskaber, der kan måle den samme effekt, fx smerte, hvilket måleredskab vil så blive foretrukket, såfremt effektmålet er afrapporteret med flere skalaer i samme studie eller på tværs af de udvalgte studier [8]. Derudover er det vigtigt at få afklaret, hvilken tidshorisont det er relevant at se på. Dette er vigtigt, da effekten af en given teknologi kan variere betydeligt over tid, og det kan således have stor betydning, om man vælger at fokusere på effekten efter seks uger, efter 12 uger eller efter et år [5].

**Mortalitet**

Mortalitet bør anvendes som effektmål, hvor det er relevant, da det er det effektmål, hvor der er mindst risiko for bias. Samlet overlevelse (eng. all cause mortality) er det foretrukne kliniske effektmål i time-to-event-analyser [7].

**Morbiditet (symptomer og bivirkninger)**

Denne kategori rummer kliniske effektmål og bivirkninger, som er relevante for patienten. Prioriteringen af angivne effektmål vil afhænge af selve sygdommen og formålet med behandlingen [7]. Almindelige typer effektmål inkluderer:

- sygelighed (fx blodprop, slagtilfælde)
- klinisk status (fx kolesteroltal, blodtryk)
- symptomer (fx smerter, kløe)
- funktionsevne (fx gangfunktion)

**Bivirkninger**

Det er vigtigt, at eventuelle bivirkninger eller utilsigtede hændelser, som afspejler sikkerheden ved den givne intervention, medtages som effektmål. Bivirkninger indsamles jævnligt som sekundære effektmål, og der er ofte variation studierne imellem, i forhold til hvordan bivirkninger rapporteres både i forhold til terminologi og detaljeringsgrad [9]. Alvorlige bivirkninger (eng. serious adverse events) forventes at forekomme relativt sjældent, og studier er som udgangspunkt ikke designet til at kunne påvise en statistisk signifikant forskel i forekomsten af bivirkninger mellem behandlingsgrupper [10]. Det skal vurderes, hvorvidt studiets design tager højde for dette og muliggør opsporing af eventuelle forskelle i forekomsten af såvel alvorlige bivirkninger som eventuelle andre relevante bivirkninger.

**Helbredsrelateret livskvalitet**

Det er velkendt, at patientens helbredsstatus også påvirker det fysiske og psykiske velbefindende. Den 'helbredsrelaterede livskvalitet' (eng. health-related quality of life, HRQoL), opfanges ikke nødvendigvis af de anvendte mortalitets- og morbiditetsmål [1]. Ved HRQoL-målinger anvender man måleinstrumenter, der indeholder brede kliniske, funktionelle og psykosociale dimensioner. Det kliniske effektmål bredes således ud til at inkludere fysiske, psykiske og sociale forhold [7]. Netop fordi HRQoL-målinger baseres på flere dimensioner relateret til patienternes sygdom og behandling, er effektmålet modtageligt over for en lang række ændringer/ eksterne faktorer. HRQoL-målinger er således ikke fyldestgørende som primære/enkeltstående effektmål, men må bedømmes sammen med andre relevante effektmål relateret til sygelighed og/eller død.

Der findes flere forskellige typer af redskaber til måling af HRQoL. Overordnet kan måleredskaberne inddeles i generiske og sygdomsspecifikke. De generiske måleredskaber dækker dimensioner, der vurderes at være vigtige for den helbredsrelaterede livskvalitet generelt, mens de sygdomsspecifikke redskaber fokuserer på dimensioner, som er påvirket af den givne helbredstilstand eller den givne population. Sygdomsspecifikke måleredskaber er generelt mere sensitive over for små ændringer i HRQoL end generiske redskaber. Derimod er generiske måleredskaber mere omfattende og vil derfor være bedre til at opfange uforudsete effekter, som ikke måles i de sygdomsspecifikke redskaber [11]. En forudsætning for, at data omkring helbredsrelateret livskvalitet kan indgå i litteraturgennemgangen, er, at det anvendte måleredskab er valideret til den patientgruppe, som redskabet er anvendt på i studiet. Eksempler på alment anvendte sygdomsspecifikke måleredskaber, som er validerede i en dansk kontekst, er EORTC QLQ-C30 til vurdering af HRQoL blandt personer med cancer og HeartQol til vurdering af HRQoL hos personer med iskæmisk hjertesygdom. Eksempler på generiske måleredskaber, som er valideret i en dansk kontekst, er EQ-5D, SF36, SF12 og 15D.

### 2.1.6 Surrogatmål

Surrogatmål er effektmål, som anvendes som erstatning for det effektmål, vi egentlig ønsker at vide noget om. Surrogatmål anvendes hyppigt inden for medicinsk forskning, primært for tidligere og mere simpelt at kunne præsentere konklusioner vedr. den afledte patientrelaterede merværdi af en given behandling. Et surrogatmål forventes at kunne prædiktere kliniske effekter baseret på epidemiologisk, patofysiologisk, terapeutisk eller anden videnskabelig evidens (12). Eksempler på surrogatmål er blodtryk som surrogatmål for hjerte-kar-sygdom eller HIV1-RNA viral load som indikation for viral suppression ved HIV-intervention [7]. En liste over surrogatmål, som ofte anvendes som substitut for patientrelaterede effektmål, ses i tabel 1.

**Tabel 1. Eksempler på patientrelaterede effektmål og surrogatmål**

| Sygdom/helbredstilstand | Patientrelaterede effektmål | Surrogatmål |
|---|---|---|
| Diabetes mellitus | Diabetiske symptomer, indlæggelse, komplikationer (kardiovaskulære, øjne, nyrer, neuropati) | Blodglukose, HbA1c |
| Hypertension | Kardiovaskulær død, myokardieinfarkt, apopleksi | Blodtryk |
| Osteoporose | Frakturer | BMD |
| KOL | Livskvalitet, eksacerbation, mortalitet. | Lungefunktion (FEV1) |
| Hjerte-kar-sygdom | Vaskulære events | Serumlipider |
| Blodprop | Blodprop | Asymptomatisk blodprop |

Forkortelser: HbA1c = langtidsblodsukker; BMD = bone mineral density (knogletæthed); FEV1 = Forced Expired Volume in the first second.

Vurdering af effekten af en teknologi alene på baggrund af surrogatmål bør altid ske med stor omtanke og forsigtighed [13], da surrogatmål kan være vildledende i forhold til vurdering af merværdi for patienterne [14]. Mange interventioner reducerer risikoen for surrogatudfald, men har ingen eller endda skadelig effekt på klinisk relevante effektmål, mens andre interventioner ikke har nogen effekt på surrogatmål, men har effekt på klinisk relevante udfald [15[. Surrogatmål bør derfor kun inddrages, såfremt der ikke findes evidens af høj kvalitet inden for det effektmål, vi ønsker at vide noget om. I dette tilfælde skal de ønskede patientrelaterede effektmål og de surrogatmål, der er associeret, beskrives [13]. Anvendelse af surrogatmål vil som beskrevet senere altid medføre overvejelse omkring nedgradering grundet indirekte evidens, jf. GRADE [3].

### 2.1.7 Kompositte effektmål

Kompositte effektmål kombinerer to eller flere udfald fx mortalitet, non-fatal myokardieinfarkt og indlæggelse i et effektmål, som samler den overordnede behandlingseffekt. Kompositte effektmål anvendes ofte, hvor den statistiske styrke (eng. statistical power) er lav, mhp. på at øge eventrate og reducere 'sample size'. Kompositte effektmål bør generelt ikke anvendes, medmindre der er en berettiget årsag til dette (fx ved sjældne sygdomme/udfald) [7].

### 2.1.8 Vægtning af effektmål

De patientrelaterede effektmål rangordnes som *kritiske*, *vigtige* og *mindre vigtige* efter, hvor vigtige de er for beslutningstagning. Denne proces pågår som udgangspunkt, inden litteratursøgningen er udført, det vil sige uafhængigt af kendskab til, hvilke effektmål der er inddraget i den publicerede litteratur. Vigtigheden af de identificerede effektmål rangeres på en skala fra 1-9 (7-9: kritiske effektmål, 4-6: vigtige effektmål og 1-3: ikke vigtige effektmål (bør derfor udelades)) for at skelne mellem de forskellige kategorier. Rangordning eller vægtning af effektmålene har ud over at hjælpe til udvælgelse af effektmål en vigtig funktion, når afvejningen mellem gavnlige og skadelige effekter skal vurderes [3]. Her bemærkes det, at inddragelse af patienter med den specifikke lidelse i arbejdsgruppen vil øge tilliden til rapporten, der vurderer værdien af en teknologi.

### 2.1.9 Udvælgelse af in- og eksklusionskriterier

I udvælgelse af evidens er det vigtigt at undgå systematiske fejl. Derfor er det essentielt, at beslutning om in- og eksklusion af studier er baseret på forhåndsbestemte kriterier. In- og eksklusionskriterier skal logisk følge det fokuserede spørgsmål, vi ønsker at besvare. Kriterierne defineres ved at tydeliggøre, hvilke populationer, interventioner, sammenligninger og effektmål vi ønsker at dække i gennemgangen (og hvilke vi ikke er interesserede i) [16].

Ud over afgrænsning på baggrund af PICO skal der også træffes beslutning om afgrænsning på baggrund af studiedesign. Afgrænsning af studiedesign medtages, da vi ønsker at basere vurderingen på den bedst tilgængelige evidens. Randomiserede, kliniske studier (RCT'er) vurderes initialt at være af høj metodisk kvalitet (på grund af det kausale element i metoden), hvor effekten af det tiltag, vi ønsker at vurdere, er sammenholdt med effekten af det relevante sammenligningsalternativ og foretrækkes som oftest som evidensgrundlag [17] Såfremt der ikke findes studier, som direkte sammenligner de tiltag, vi er interesserede i, kan det være nødvendigt at søge efter indirekte sammenligninger via RCT-studier. I tilfælde, hvor der ikke findes RCT'er, kan det overvejes at inkludere andre studiedesigns (4). Her bemærkes det, at der ofte er en fejlagtig antagelse i 'evidens-baseret medicin' af, at (dårlige) RCT-studier er bedre end fx (gode) kohortestudier hvilket ikke er korrekt [18]. Det sker ofte, at tilliden til evidensen er lige god for såvel RCT'er som for kohortestudier [19].

### 2.1.10 Udarbejdelse af projektprotokol

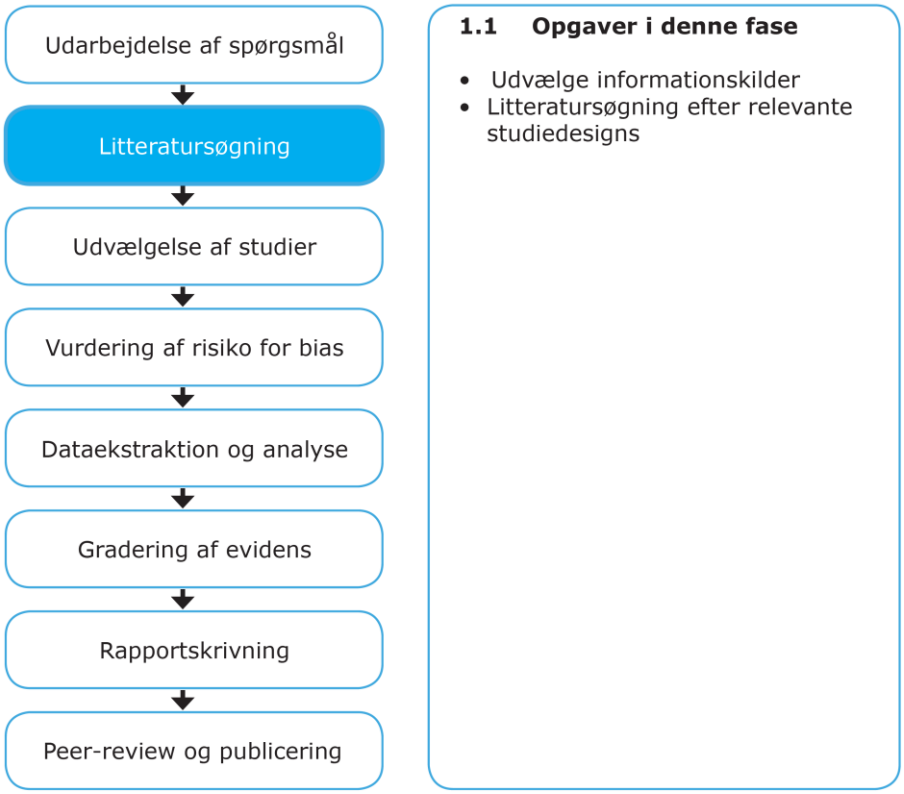
Forberedelsesfasen resulterer i udarbejdelse af en projektprotokol. Projektprotokollen skal indeholde en overordnet beskrivelse af baggrunden for litteraturgennemgangen samt en beskrivelse af, hvilke spørgsmål den givne gennemgang skal besvare, herunder afgrænsning af PICO. Derudover skal protokollen indeholde en beskrivelse af, hvordan litteraturgennemgangen skal gennemføres, herunder beskrivelse af strategi for søgning og udvælgelse af litteratur. Formålet med udarbejdelse og publicering af en projektprotokol er at facilitere en stringent og transparent proces for gennemførelsen af litteraturgennemgange [20].

**Projektprotokollen skal indeholde en beskrivelse af følgende**

- Sygdommen/tilstanden og aktuelle kliniske håndtering af tilstanden.
- Tekniske karakteristika ved den nye teknologi/tiltag som skal vurderes.
- Domænespecifikke spørgsmål som gennemgangen skal besvare inkl. PICO.
- Søgestrategi, søgetermer og informationskilder.
- In- og eksklusionskriterier for studier.
- Hvordan studierne skal kvalitetsvurderes.
- Hvordan data ekstraheres, og hvilke data der ekstraheres.
- Hvordan informationen opsummeres (syntetiseres).
- Tidsplan for gennemgangen.

## 2.2 Informationssøgning

**Figur 2. Opgaver i informationssøgningsfasen***




*Figuren er bearbejdet på baggrund af figurer i 'Håndbok for Nasjonalt kunnskapssenter for helsetjenesten desember 2015' – 'Slik oppsummerer vi forskning'.

### 2.2.1 Vejledning for litteratursøgning

Planlægning og udarbejdelse af en effektiv litteratursøgning kræver særlige kompetencer, og det vil derfor ofte være hensigtsmæssigt at involvere en litteratursøgningsekspert tidligt i processen. Derudover vil det i mange tilfælde være givtigt at udvikle søgestrategien i tæt samarbejde med kliniske eksperter inden for det område, man undersøger. Kliniske eksperter og søgeeksperter har viden og kompetencer, som er afgørende for at udarbejde en effektiv søgestrategi. De kliniske eksperter kender til det faglige område, fagtermer og centrale studier inden for området, mens søgeeksperterne kan hjælpe med at opstille en brugbar søgestrategi. Nedenfor følger en kort gennemgang af de forskellige faser i en systematisk litteratursøgning. Uanset om man selv gennemfører søgningen eller får en søgeekspert til at gennemføre søgningen, er det vigtigt at kende til de forskellige faser. Hvis man selv skal gennemføre den systematiske litteratursøgning og ikke har meget erfaring med dette, anbefales det, at man søger yderligere information om litteratursøgning ud over denne gennemgang.

#### 2.2.1.1 Søgestrategi

Det første led i udarbejdelsen af en effektiv søgestrategi er udarbejdelsen af fokuserede spørgsmål, som udspringer af projektets overordnede problemstilling. Det vil sige, at man udarbejder et eller flere spørgsmål, der specificerer, hvad der ønskes søgt efter i de udvalgte databaser. En god og velafprøvet måde at udarbejde disse spørgsmål på er ved at benytte PICO-tilgangen, der er udarbejdet til at strukturere kliniske spørgsmål og søgeprocesser, men den kan også benyttes ved andre former for problemstillinger, hvor man kan tilpasse kategorierne, så de passer til det valgte område.

I udformningen af de fokuserede spørgsmål drejer det sig således om at identificere hvilken patientgruppe/population/problemområde og hvilken intervention man ønsker at undersøge. Derudover skal man have identificeret, hvilken anden teknologi, medicin, organisering osv. man ønsker at sammenligne med (komparator), og som er relevant for den valgte problemstilling. Det frarådes som udgangspunkt at søge på outcome, medmindre reviewet specifikt omhandler et bestemt outcome, fx mortalitet. Det er vigtigt også at være struktureret i denne del af litteratursøgningen, da dette arbejde lægger grunden for gennemførelsen af en vellykket litteratursøgning. Nedenfor er vist et eksempel på en struktureret identifikation og opstilling af et fokuseret spørgsmål, der kan danne baggrund for en litteratursøgning.

| **Problemstillingen bør formuleres som et spørgsmål**<br>Spørgsmålet skal bestå af følgende 4 (eventuelt 3) dele: |
|---|
| 1. **Patient/problem.** Hvilke patienter/tilstand/sygdom drejer det sig om?<br><br>Voksne patienter med mekanisk hjerteklap<br><br>2. **Intervention.** Hvilken intervention/eksposition drejer det sig om?<br><br>Selvstyret antikoagulationsbehandling<br><br>3. **Komparator.** Hvad sammenlignes interventionen med?<br><br>Konventionel antikoagulationsbehandling (fremmøde på hospitalet til tjek)<br><br>4. **Outcome.** Hvilke effekter/udfald er af interesse?<br><br>Livskvalitet, blødninger, tromboemboliske komplikationer og dødelighed. |
| **Skriv hele spørgsmålet her:** |
| Hvilke effekter har selvstyret antikoagulationsbehandling for voksne patienter med mekanisk hjerteklap sammenlignet med konventionel antikoagulationsbehandling, set i forhold til terapeutisk INR-interval, livskvalitet, blødninger, tromboemboliske komplikationer og dødelighed? |
| |

| **Hvilke søgeord er aktuelle for at dække problemstillingen?** | | | | |
|---|---|---|---|---|
| Brug engelske ord, og vær påpasselig med at få alle synonymer med (herunder også velkendte stavefejlstyper i litteraturen). Det er en fordel at dele søgeordene op efter, hvad der gælder/beskriver patienten, interventionen/eksposition en, sammenligningen og udfaldet. | | | | |
| | **AND** | | | |
| | **Patient/Problem** | **Intervention** | **Comparator** | **Outcome** |
| **OR** | Mechanical cardiac valve<br><br>Mechanical heart valve<br><br>Valve prosthesis | Self administration<br><br>Self-management-anticoagulation<br><br>Self-monitoring<br><br>Self-administration<br><br>Patient self-management | Ambulatory care<br><br>Outpatient clinics<br><br>Monitoring ambulatory<br><br>Ambulatory care facilities | Therapeutic INR intervals<br><br>Quality of life<br><br>Haemorrhage<br><br>Thrombosis<br><br>Mortality |

Med udgangspunkt i de fokuserede spørgsmål udarbejdes en søgestrategi. Det er her vigtigt at være opmærksom på, at de forskellige databaser kræver hver sin søgestrategi. Søgninger i bibliografiske databaser baseres på en detaljeret søgestrategi, hvor det skrives ned, hvordan der søges i den pågældende database: hvilke søgetermer der anvendes, og hvordan disse termer kombineres for at gøre søgningen så præcis og dækkende som muligt. Søgestrategi og udvælgelse af søgetermer udarbejdes gerne i samarbejde mellem søgespecialisten, kliniske eksperter og projektudfører. Som hovedregel er en god struktur for en systematisk søgning at opbygge den ud fra de forskellige elementer, man har identificeret i det fokuserede spørgsmål. Hvis man benytter PICO-modellen, har man således fire elementer, der kan kombineres på forskellige måder. Det kræver, at man prøver sig frem for at finde den bedste kombination og søgestrategi, og ofte at der ikke søges på baggrund af alle fire elementer, men fx på P og I.

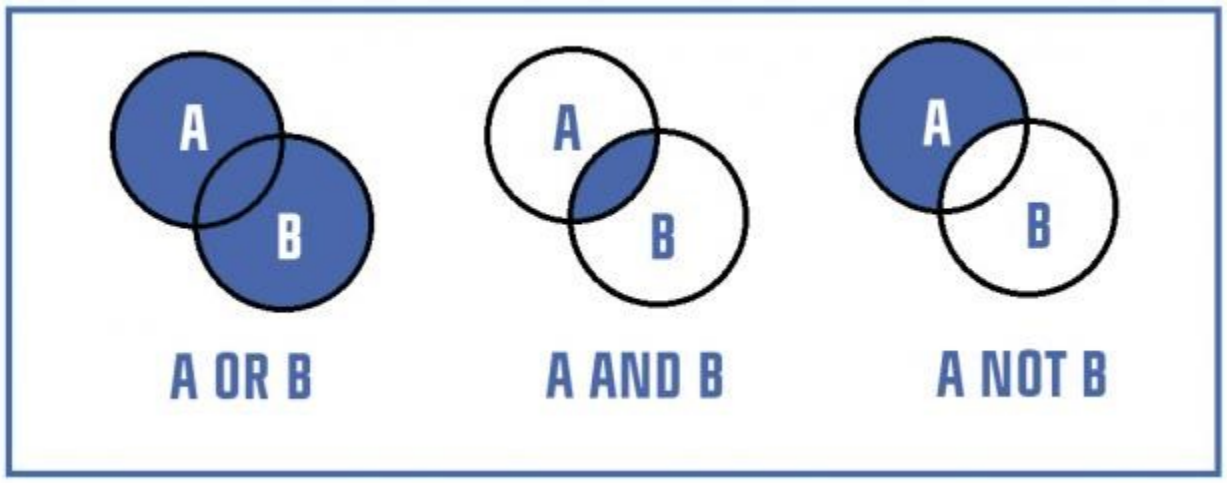
Når man søger i de valgte databaser, bør man som udgangspunkt anvende kontrollerede emneord for at sikre, at alle relevante studier indfanges i søgningen. Fordelen ved at anvende kontrollerede emneord er, at man også fanger studier, som anvender andre termer for de emner, man søger på, end dem man selv anvender, da emneord indekserer alle studier, der omhandler emnet, uanset hvilke ord der benyttes i studiet til at beskrive præparatet eller tilstanden. De kontrollerede emneord er hierarkisk opbygget med overordnede, underordnede og sideordnede termer, hvilket gør det muligt at søge på den valgte term, samtidig med at man også søger på studier indeholdende underordnede termer til det valgte emneord. Derudover vil brugen af kontrollerede emneord også hjælpe til at fokusere litteratursøgningen, da artikler

kun bliver indekseret med et emneord, hvis indholdet i artiklen i væsentlig grad omhandler dette emneord.

Hvis man ønsker at gøre søgningen mere bred og være sikker på at få så meget litteratur med som muligt, eller hvis der ikke fremkommer tilstrækkelig mængde af relevante hits ved anvendelse af kontrollerede emneord, er det en mulighed at inkludere fritekstord og synonymer i søgningen. Fritekstord er ord man søger frit i databaserne uden at knytte dem til et emneord. Søgningen kan ligeledes udvides ved at trunkere de valgte søgeord, således at forskellige bøjninger og sammensætninger af et ord kommer med (søg fx på child* for at finde både children, childhood, child, childcare og children's).

Når man har defineret søgeordene, skal det besluttes, hvordan termerne skal kombineres. Her benytter man som udgangspunkt de tre boolske operatorer AND, OR eller NOT. Hovedreglen er, at alle søgeord, der dækker der samme område, fx population. skal kombineres med OR, mens de forskellige områder efterfølgende kombineres med AND. For et eksempel, se nederst i figur 3 (eksemplet med det opstillede spørgsmål ovenfor). Figur 3 viser, hvilket udfald de tre boelske operatorer giver. Reelt set frarådes anvendelse af 'NOT' i sin 'boolean', da brugen af denne operator ofte medfører eksklusion af relevante studier.

**Figur 3. Boolske operatorer**




### 2.2.1.2 In- og eksklusionskriterier

Ved at fastsætte inklusionen og eksklusions kriterier inden man foretager sin søgning, kan man gøre søgningen mere præcis og derved medvirke til at reducere i antallet af referencer, der i sidste ende skal gennemgås. Eksempler på in- og eksklusionskriterier kan ses i følgende liste.

- Er det kun bestemte studiedesigns, der skal inkluderes fx metaanalyser, reviews, RCT, kohorte-studier osv?
- Er det kun bestemte aldersgrupper, der skal inkluderes, fx børn, 40-60 årige, unge osv?
- Er det kun studier fra en bestemt tidsperiode, der skal inkluderes, fx studier fra 2000 og frem, studier fra de sidste 10 år osv.?
- Er der kun studier fra bestemte lande der skal inkluderes?
- Skal kun mænd eller kvinder inkluderes?
- Er det kun artikler skrevet på bestemte sprog der skal inkluderes?

Ud over disse eksempler kan der være andre relevante kriterier, der gælder specifikt for den pågældende søgning.

#### 2.2.1.3 Udarbejdelse af søgeprotokol

Et vigtigt element i planlægningen af litteratursøgningen er udarbejdelse af en søgeprotokol. En søgeprotokol har til formål at skabe overblik og transparens i processen med at indsamle evidens. Ligeledes er søgeprotokollen med til at sikre konsistens ved opfølgning eller gentagelse af søgningen.

Søgeprotokollen skal indeholde følgende elementer:

- Baggrund og problemstilling
- Fokuserede spørgsmål
- In- og eksklusionskriterier (defineret på forhånd)
- Informationskilder (databaser, internetsider, registre etc.)
- Søgestrategi for hver enkelt informationskilde
- Eventuelle restriktioner (fx publikationstype)
- Strategi for gennemgang og udvælgelse af den fundne litteratur

#### 2.2.1.4 Databaser og andre informationskilder

Der findes mange forskellige kilder til informationssøgning, og hvilke kilder, der er relevante for en given problemstilling, afhænger af flere forskellige faktorer. Blandt andet hvilken type litteraturgennemgang skal man lave, hvilke type spørgsmål skal besvares (effekt, diagnose, organisering) og nogle gange også tidshorisonten for opgaven. I tabel 2 ses eksempler på forskellige kilder og til sidst henvises til andre oversigter, der er mere udførlige.

**Tabel 2. Eksempler på kilder til identifikation af studier**

| **Eksempler på kilder til identifikation af sekundære studier** |
|---|
| Cochrane Database of Systematic Reviews |
| The HTA Database: www.crd.york.ac.uk/CRDWeb/ |
| Guidelines International Network (G-I-N): www.g-i-n.net |
| National Guidelines Clearinghouse : www.guideline.gov |
| EMBASE |
| PubMed |
| **Eksempler på kilder til identifikation af primære studier** |
| PubMed |
| EMBASE |
| Cinahl |
| PsycINFO (primært patient) |
| Sociological Abstract (primært patient) |
| Social Sciences Citation Index (primært patient) |
| **Eksemplar på kilder til identifikation af grå litteratur** |
| Grey Matters (oversigt over kilder): www.cadth.ca/resources/finding-evidence/grey-matters |
| Faglige selskabers hjemmesider både nationalt og internationalt |
| MTV-organisationers hjemmesider både nationalt og internationalt |
| Generel søgning på nettet |

For mere information om kilder til informationssøgning kan følgende publikationer anbefales:

- Metodehåndbog for medicinsk teknologivurdering [1]
- METODEHÅNDBOGEN. Model for udarbejdelse af nationale kliniske retningslinjer [5]
- Slik opsummerer vi forskning [16].

#### 2.2.1.5 Evaluering og eventuel revidering af søgestrategi

Søgestrategien udvikles og tilpasses ved at udføre prøvesøgninger. Hvis der kommer for mange irrelevante hits, eller hvis relevante referencer ikke fremkommer i søgningen, bør strategien revideres ved at tilføje eller fjerne emne- og fritekstord. Søgestrategien kan blandt andet evalueres ved at krydstjekke, om udvalgte relevante artikler indfanges i den planlagte søgestrategi. Fremkommer den udvalgte artikel ikke ved søgningen, kan dette skyldes, at artiklen er for ny, eller at tidsskriftet ikke er med i den pågældende database. Findes artiklen

derimod i databasen (men ikke i søgeresultatet), indikerer dette, at søgestrategien ikke er tilstrækkeligt præcis. Det vil derfor være nødvendigt at revidere søgestrategien. En god metode til at præcisere sin søgestrategi er at inkludere de emneord i søgestrategien, som den relevante artikel, der ikke fremkom i søgningen, er blevet tildelt ved registrering i databasen.

#### 2.2.1.6 Dokumentation af litteratursøgning

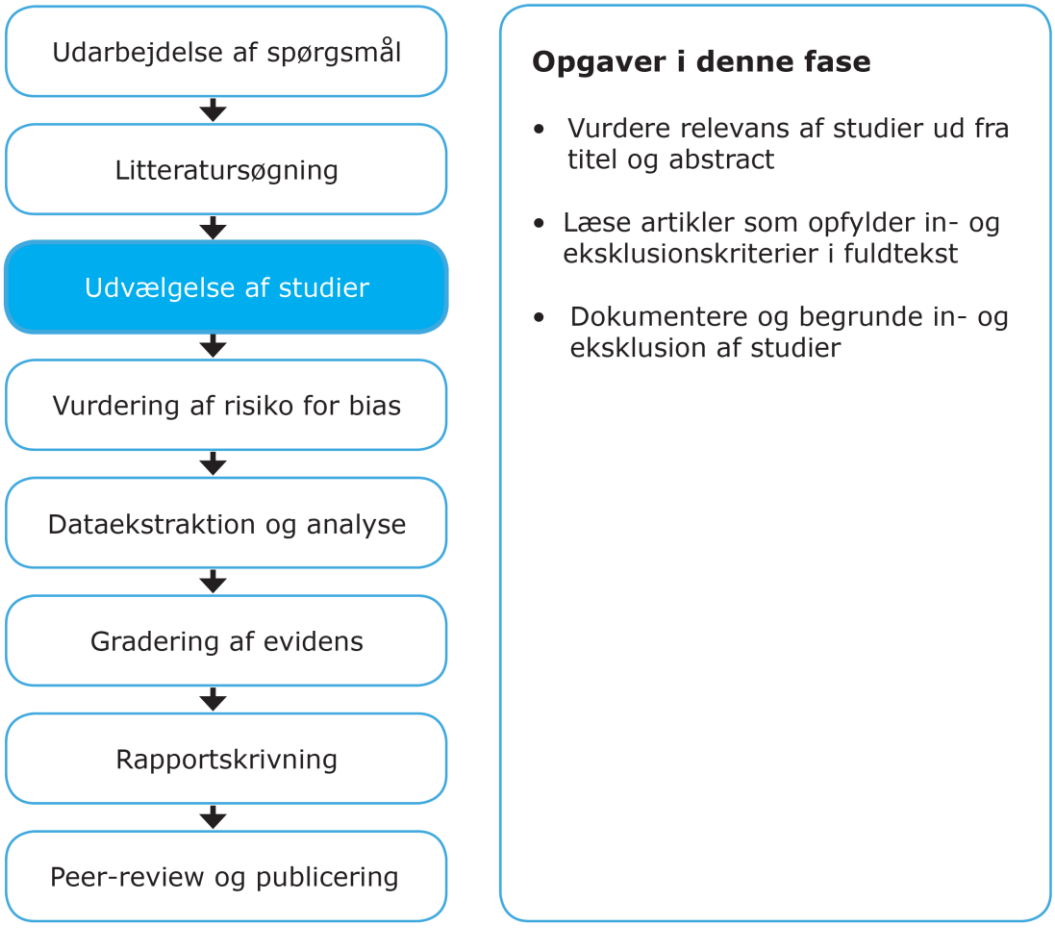
Som udgangspunkt skal man altid beskrive og dokumentere de søgninger, der gennemføres. Hvis man udarbejder en søgeprotokol, er man godt på vej i forhold til at dokumentere de gennemførte søgninger. Derudover vil det være hensigtsmæssigt at inkludere et flowdiagram, der dokumenterer, hvor mange referencer der blev identificeret, og hvilke der blev sorteret fra, jf. PRISMA-flowdiagram [21]. Den specifikke søgehistorie fra de enkelte databaser skal vedlægges som bilag, eventuelt via en søgeprotokol. Dokumentationen for søgningen skal være så detaljeret, at man ved at følge beskrivelsen kan gentage søgningen i de udvalgte databaser.

#### 2.2.1.7 Bearbejdning af referencer ved hjælp af referencehåndteringsværktøj

Den letteste måde at håndtere de identificerede referencer på er at anvende et referencehåndteringsværktøj. Ved vurdering af lægemidler anvendes fx referenceprogrammet EndNote. EndNote er et webbaseret program, hvor man kan opbygge emneinddelte databaser over litteraturreferencer. Referencer fra stort set alle databaser kan importeres direkte til EndNote. Derudover er der mulighed for manuel indtastning af referencer. EndNote er velegnet til at skabe overblik over de identificerede referencer fra søgningerne. Når referencerne er importeret fra de forskellige databaser, kan dubletter (dvs. referencer som findes i mere end én database) fjernes automatisk, og sortering af referencerne kan påbegyndes. EndNote anvendes ligeledes til indsættelse af kilder i baggrundsnotatet samt ved udarbejdelse af automatiske litteraturlister i det ønskede format.

## 2.3 Udvælgelse af studier

**Figur 4. Opgaver ved udvælgelse af studier***




*Figuren er bearbejdet på baggrund af figurer i 'Håndbok for Nasjonalt kunnskapssenter for helsetjenesten desember 2015' – 'Slik oppsummerer vi forskning'.

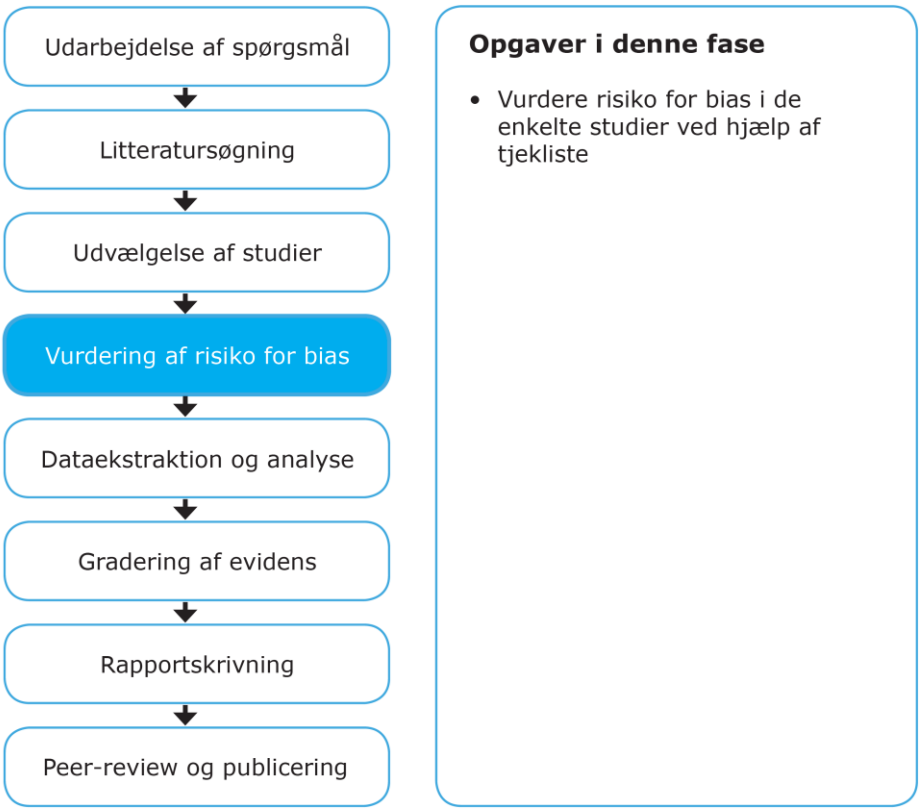
Når søgningen i databaserne er gennemført og dubletter er fjernet, kan udvælgelse af relevante studier påbegyndes. Udvælgelse foretages på baggrund af de eksplicitte in- og eksklusionskriterier. Det anbefales, at relevansen af de identificerede referencer vurderes uafhængigt af minimum to personer fra arbejdsgruppen [1]. Første gennemgang af referencer foregår som oftest på overskrift- og abstractniveau. Referencer, som vurderes at være irrelevante, frasorteres. Litteratur læses i fuldtekst når:

- begge projektgruppemedlemmer vurderer, at referencen kan være relevant.
- projektgruppemedlemmerne er uenige om eller usikre på, hvorvidt referencen er relevant.

Den udvalgte litteratur læses nu i fuldtekst af mindst to projektgruppemedlemmer, og beslutning om in- og eksklusion træffes. Ekskluderede studier samt årsag til eksklusion noteres [1].

## 2.4 Vurdering af risiko for bias i primære studier

**Figur 5. Opgaver ved udvælgelse af studier***




*Figuren er bearbejdet på baggrund af figurer i 'Håndbok for Nasjonalt kunnskapssenter for helsetjenesten desember 2015' – 'Slik oppsummerer vi forskning'.

Vurdering af de inkluderede studiers interne validitet (tidligere ofte omtalt som metodiske kvalitet) er vigtig med henblik på at kunne vurdere, i hvor høj grad vi kan stole på de konklusioner, som præsenteres i studierne [22]. Der findes forskellige tjeklister, som kan anvendes til kvalitetsvurdering af studier [23]. Tjeklisterne består af en række kriterier, som anvendes til en kritisk vurdering af kendte kilder til systematiske fejl også kaldet bias. Det er god forskningspraksis at to projektgruppemedlemmer vurderer risikoen for bias uafhængigt af hinanden [16].

### 2.4.1 Vurdering af systematiske reviews

Tjeklisten ROBIS (Risk of Bias In Systematic Reviews[2]) anbefales til vurdering af kvaliteten af systematiske reviews. ROBIS består af tre faser. I fase 1 vurderes relevansen af reviewet i forhold til det emne, man ønsker at vide noget om – dvs. i hvor høj grad reviewets forsknings-

[2] https://www.bristol.ac.uk/population-health-sciences/projects/robis/

spørgsmål stemmer overens med det spørgsmål, som ønskes besvaret. I fase 2 vurderes risikoen for bias i fire domæner, som dækker følgende nøgleelementer i review-processen:

1. Egnethedskriterier for de inkluderede studier
2. Identifikation og udvælgelse af studier
3. Dataudtræk og kvalitetsvurdering af studier
4. Syntese og resultater

I domæne 1 (egnethedskriterier) foretages en vurdering af, om formål samt in-og eksklusionskriterier for reviewet var defineret på forhånd, og hvorvidt de valgte kriterier for litteratursøgningen var hensigtsmæssige. Domæne 2 (identifikation og udvælgelse af studier) omfatter en vurdering af, om litteratursøgningen har været tilstrækkeligt omfattende i forhold til at få identificeret alle relevante referencer, samt om udvælgelsen af studier er foretaget i overensstemmelse med god forskningsskik (kvalitetssikring). I domæne 3 (dataudtræk og kvalitetsvurdering) vurderes det, om dataekstraktionen er udført systematisk og med passende grad af kvalitetssikring. Ydermere foretages en formel kvalitetsvurdering af de inkluderede studier. Domæne 4 (syntese og resultater) indeholder en vurdering af, hvorvidt syntesen er udarbejdet på en hensigtsmæssig måde. Såfremt syntesen indeholder en kvantitativ sammenfatning af resultaterne (metaanalyse), omfatter bedømmelsen dels en vurdering af, hvorvidt det var hensigtsmæssigt at sammenfatte resultaterne statistisk og dels en vurdering af de statistiske analyser, som er anvendt. I fase 3 opsummeres vurderingerne fra fase 2, og den samlede risiko for bias fastsættes.

I opsummeringen af systematiske reviews skal der ikke foretages en vurdering af de primære studier, som indgår i oversigtsartiklerne, idet vi i denne sammenhæng tager udgangspunkt i review-forfatternes vurderinger af primærstudierne.

### 2.4.2 Vurdering af primærstudier

Der findes forskellige typer af tjeklister til vurdering af bias i primære studier, afhængig af om studiet er et RCT-, kohorte-, case-kontrol-, diagnostisk eller prognostisk studie.

I tabel 3 ses en oversigt over de tjeklister, der anvendes ved vurdering af primærstudier.

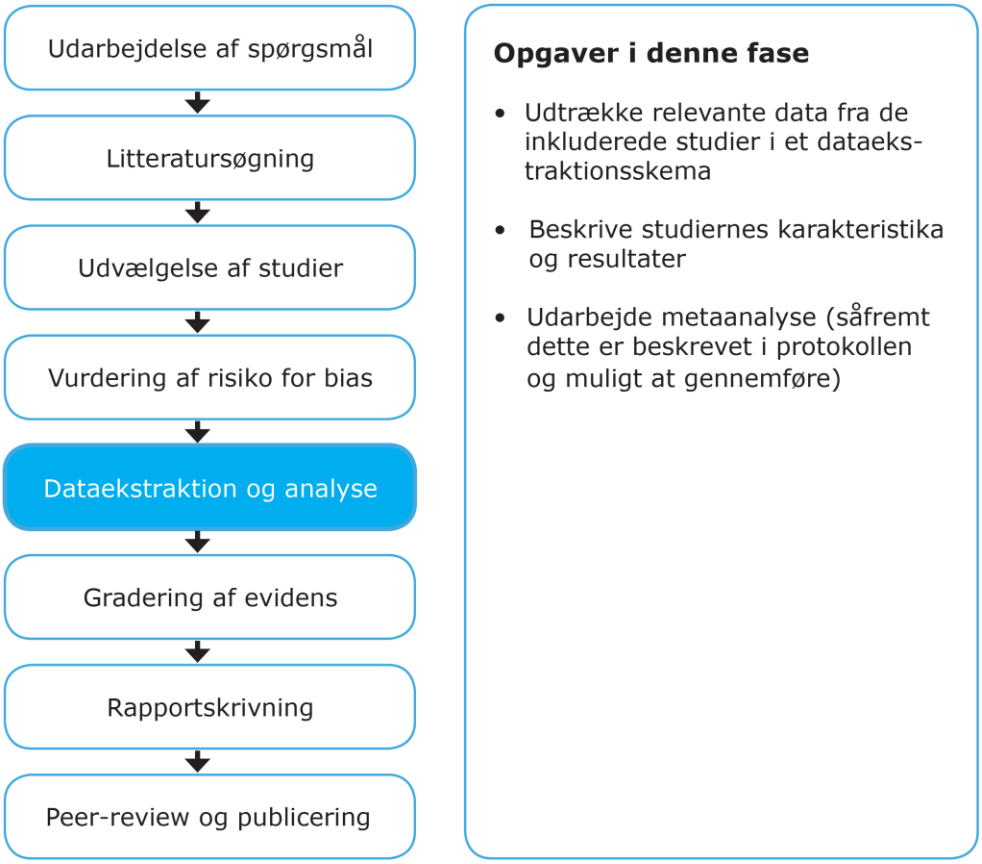
**Tabel 3. Tjeklister som anvendes ved vurdering af forskellige typer af studier**

| Studiedesign | Tjekliste | Reference |
|---|---|---|
| RCT | Cochrane Risk of Bias Tool | http://handbook.cochrane.org/ |
| Observationelle studier (kohorte, case-kontrol) | ROBINS-I tool (Risk of bias in non-randomized studies - of interventions) | [23] |
| Diagnostiske studier | QUADAS-2 | |
| Prognostiske studier | QUIPS | |

Efter at de inkluderede primære studier er gennemgået ved hjælp af tjeklister, foretages en samlet vurdering af risikoen for bias på den samlede evidens - på tværs af studier - for hvert effektmål. Denne proces er beskrevet i afsnit 2.6 "Gradering af estimater (GRADE)".

## 2.5 Dataekstraktion og sammenfatning af resultater

**Figur 6. Opgaver ved dataekstraktion og sammenfatning af resultater***




*Figuren er bearbejdet på baggrund af figurer i 'Håndbok for Nasjonalt kunnskapssenter for helsetjenesten desember 2015' – 'Slik oppsummerer vi forskning'.

I dette afsnit beskrives den metodiske tilgang til udtræk og analyse af relevant information fra de inkluderede studier. Formålet med dataudtræk og analyse er på tværs af studier at få et overblik over den samlede effekt af en given intervention.

### 2.5.1 Dataekstraktion

Ved dataekstraktion indsamles relevante oplysninger fra inkluderede studier. Processen bør ske systematisk, konsekvent og stringent for hvert enkelt studie, og samme typer af oplysninger bør indsamles for de enkelte studier. Det er hensigtsmæssigt at anvende et dataeks-traktionsskema (tabel 4). Dataekstraktionen bør udføres uafhængigt af to personer [16]. Eventuelle uenigheder løses ved diskussion eller ved at inddrage tredjepart. Processen bør dokumenteres, og det er vigtigt, at beskrivelsen af studierne ikke efterlader tvivl om karakteristika for patientgruppe, intervention og sammenligningsgrundlag, herunder kontekst. En systematisk tilgang til præsentation af dataekstraktionen er således vigtig (tabel 4).

**Tabel 4. Eksempel på tabelstruktur**

| **Forfattere<br>År<br>Land<br><br>Reference-nummer og publika-tionsdato** | **Formål, studie-design, statistik** | **Population, patientka-rakteristika** | **Intervention/ sammenligning** | **Followup-periode<br>Dropout<br>Målemetode<br>Effektmål** | **Resultater** | **Studiets kvalitet jf. tjeklister<br><br>Kommen-tarer** |
|---|---|---|---|---|---|---|
| | Formål (RCT, CT, kohorte, case-control etc.)<br>fx ITT, per protokol | In-/eksklu-sionskriterier, setting, antal ved baseline i behandlings-grupper, køn og alder, diagnose, alvorligheds-grad, komorbiditet | Intervention (dosis, interval, varighed)<br><br>Kontrol (aktiv, placebo, sædvanlig behandling, etc.) | Dropout (%) (Fra baseline til followup, eller fra afslutning af intervention til followup) | HR, RR, OR, p-værdi, konfidens-intervaller på estimat, sensitivitet, specificitet, absolut forskel etc. | |

CT = Controlled trial; HR = Hazard ratio; OR = Odds ratio
RCT = Randomised controlled trial; RR = Risk ratio

## 2.5.2 Beskrivelse af studiers resultater

Estimater inden for de enkelte effektmål præsenteres som et samlet estimat på baggrund af metaanalyse, hvor dette er statistisk muligt. Metaanalyseresultaterne præsenteres med så-kaldte 'forest plots', der illustrerer enkeltstudier samt evidenssyntesen, der leder til samlede estimater og heterogenitetsstatistik etc. Hvis datagrundlaget ikke egner sig til metaanalyse på grund af klinisk eller effektmålsheterogenitet mellem studierne (metodemæssigt eller klinisk, jf. PICO), eller hvis det ikke er muligt at gennemføre metaanalyse (enkeltstudier, andre årsa-ger), præsenteres resultater for effektmål på baggrund af enkeltstående studier i tekstform. Resultater fra studierne kan analyseres i programmet Review Manager, hvori data fra primære studier indsamles, og metaanalyser kan gennemføres.

### 2.5.2.1 Typer af data

Effektmål i de inkluderede studier kan være kontinuerte (fx blodtryk eller HbA1c), binære (fx død eller sygdomsfri efter seks måneder), ordinale/kategoriske (fx Likert-skala) eller 'count'-data (fx antal hospitalsindlæggelser). Heraf er de to førstnævnte langt de hyppigst anvendte [24,25].

### 2.5.3 Fremstilling af estimater

Binære effektmål: Effektmål i de inkluderede studier kan være udtrykt som relative (Risk Ratio (RR) eller Odds Ratio (OR)), eller som absolutte (Risk Difference (RD)) - som nemt fortolkes som 'number needed to treat' (NNT). Ideelt set bør man foretage analyserne som Risk Ratio (RR) og fortolke dem via en konvertering til et absolut mål (Risk Difference (RD)). Absolutte mål er generelt nyttige for klinikere, da de giver en mere realistisk kvantificering af behandlingseffekten end relative mål. Derimod er generaliserbarheden af absolutte mål til andre populationer begrænset. Relative mål har den fordel, at de som regel er stabile (dvs. mere homogenitet) på tværs af studiepopulationer med forskellige baselineværdier og er derfor nyttige, når resultater fra forskellige studier skal kombineres i en metaanalyse [26] Relative værdier har til gengæld den ulempe, at de ikke reflekterer patienternes baselineværdier i forhold til det givne effektmål [7].

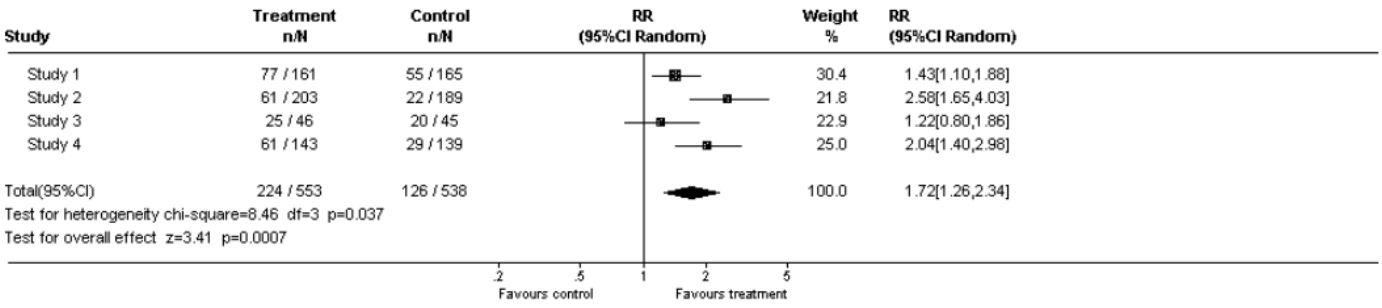
### 2.5.4 Time-to-event-data

Time-to-event-data benyttes, når man er interesseret i den tid, der går, før en hændelse optræder. De er kendt som overlevelsesdata i statistik, og død er ofte den hændelse, der måles på. Den mest hensigtsmæssige måde at sammenfatte time-to-event-data på, er at bruge overlevelsesanalyser og udtrykke effekten af interventionen som en Hazard Ratio (HR) [15].

### 2.5.5 Metaanalyse

Ved udarbejdelse af metaanalyser kan det ofte være hensigtsmæssigt at inddrage statistisk bistand mhp. udarbejdelse og fortolkning af det statiske grundlag og udarbejdelsen af 'forest plots'.

**Figur 7. Forest-plot**




Forskellige studier præsenterer ofte data på forskellige måder med brug af forskellige skalaer samt forskel i måletidspunkt o.a. De metodemæssige og kliniske forskelle, der optræder mellem studierne, skal vurderes i forhold til, om datagrundlaget er tilstrækkeligt ensartet til at

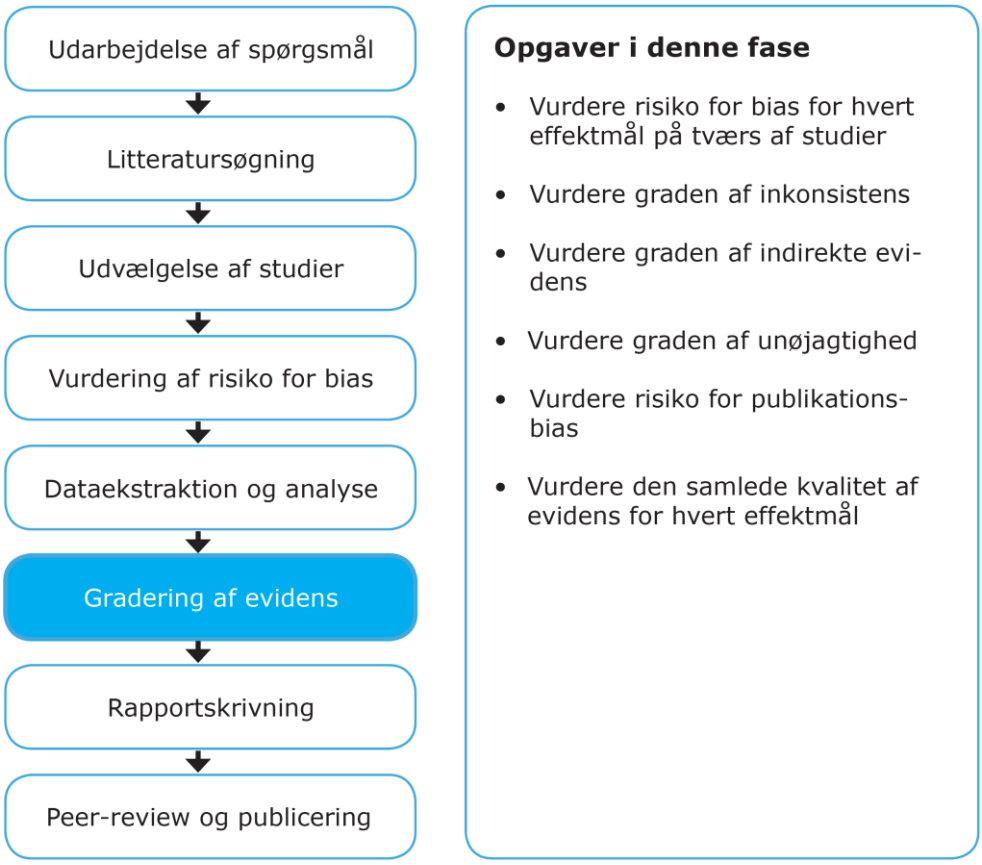
kunne kategoriseres og beskrives samlet samt i forhold til, hvilke data, der kan indgå i en metaanalyse [16].

Metaanalysen tager udgangspunkt i enkeltstudiers effektestimater og 95 % sikkerhedsintervaller (effekt ±1.96×standard error). I selve analysen foretages en vægtning af disse estimater, primært baseret på antal hændelser og deltagere i hvert enkelt studie, og et samlet effektestimat beregnes dernæst for hvert effektmål. Med det samlede estimat etableres et konfidensinterval. Et 'forest plot' er den grafiske fremstilling af analysen hvor, 'diamanten' nederst (se figur 2) viser det samlede effektestimat (og konfidensinterval) fra metaanalysen, mens firkanter og vandrette linjer viser effektestimater og konfidensintervaller fra de enkelte studier. Den vertikale linje er 'the line of no difference', og er 'diamanten' placeret hen over denne linje, kan der altså ikke påvises forskel mellem grupperne. I analysen benyttes 'random effects'-metoden til forskel fra 'fixed effects'-metoden. Denne tilgang tager højde for den forventede heterogenitet mellem studierne [15].

Homogeniteten (dvs. det modsatte af heterogenitet) mellem studierne kan vurderes i form af en Q-test som fortolkes ved $I^2$–indeks [15]. $I^2$–indekset repræsenterer procentdelen af den totale variation mellem studierne, som kan tillægges inkonsistens (statistisk heterogenitet), og som altså ikke er tilfældig variation [15].

## 2.6 Gradering af estimater (GRADE)

**Figur 8. Opgaver ved gradering af estimater***




*Figuren er bearbejdet på baggrund af figurer i 'Håndbok for Nasjonalt kunnskapssenter for helsetjenesten desember 2015' – 'Slik oppsummerer vi forskning'.

Når den systematiske gennemgang af den inkluderede primære litteratur er gennemgået, foretages en gradering af kvaliteten af evidensen. Kvaliteten af evidensen vurderes ved hjælp af GRADE (Grading of Recommendations Assessment, Development and Evaluation system), som er en anerkendt international tilgang til systematisk og transparent vurdering af evidens. Den er populært kaldet 'evidens-baseret medicin version 2.0', da den har færre svagheder, flere tilhængere (fx WHO og Cochrane-samarbejdet) og er opfundet af dem, der også startede EBM-bølgen. Ved besvarelse af undersøgelsesspørgsmål benyttes GRADE således til vurdering af evidensgrundlaget af den kvantitative litteratur for hvert enkelt af de vigtige effektmål på tværs af de inkluderede studier. Evidensvurdering og resultater på tværs af studierne samles og præsenteres i evidensprofiler. Overordnet kan man sige, at GRADE er en transparent og struktureret proces for udvikling og præsentation af den samlede evidens.

I det følgende gennemgås GRADE's tilgang til gradering af evidens trin for trin. For yderligere information om GRADE henvises til The Grade Working Group's hjemmeside

(http://www.gradeworkinggroup.org/), hvor der blandt andet findes links til en artikelserie om GRADE-metoden.

### 2.6.1 Vurdering af kvaliteten af evidensen for det enkelte effektmål

Kvaliteten af evidensen graderes i fire niveauer: Høj, moderat, lav eller meget lav. Kvaliteten af evidensen for de enkelte effektmål fastsættes med udgangspunkt i studiernes design. RCT-studier starter ved høj tillid til evidensen (på grund af den forventede kausalitet i designet), mens estimater fra observationelle studier som udgangspunkt giver os en lav tillid til evidensen [27]. I tabel 6 ses en oversigt over betydningen af de fire evidenskategorier.

Der er fem kriterier, som kan føre til nedgradering og tre faktorer, som kan føre til opgradering af tilliden til evidensen for det enkelte effektmål. De fem nedgraderingskriterier omfatter: Risiko for bias, inkonsistens, indirekte evidens, unøjagtighed og publikationsbias [28]. En opsummering af GRADE's tilgang til gradering af evidens ses i tabel 5.

**Tabel 5. Ned- og opgradering af evidens i GRADE**

| Studiedesign | Initial kvalitet af evidens | Lavere hvis | Højere hvis | Kvaliteten af den samlede evidens |
|---|---|---|---|---|
| RCT | Høj → | **Risiko for bias**<br>-1 alvorlig<br>-2 meget alvorlig | **Stor effekt**<br>+1 stor<br>+2 meget stor | Høj ++++<br>Moderat +++ |
| Observationelt | Lav → | **Inkonsistens**<br>-1 alvorlig<br>-2 meget alvorlig<br>**Indirekte evidens**<br>-1 alvorlig<br>-2 meget alvorlig<br>**Unøjagtighed**<br>-1 alvorlig<br>-2 meget alvorlig<br>**Publikationsbias**<br>-1 sandsynlig<br>-2 meget sandsynlig | **Dosis-respons**<br>+1 evidens for gradient<br>**All plausibel residual-konfounding**<br>+1 ville reducere den påviste effekt | Lav ++<br>Meget lav + |

#### 2.6.1.1 Nedgradering af evidens

Nedgradering fortages for hvert effektmål på tværs af de studier, der besvarer det fokuserede spørgsmål. Er problemet af mindre alvorlig karakter, nedgraderes med ét niveau. Er problemet af mere alvorlig karakter, nedgraderes to niveauer. I tabel 6 vises en oversigt over de fem kriterier, som kan føre til nedgradering.

**Tabel 6. Kriterier for nedgradering af tilliden til evidens**

| **Kriterier for nedgradering af evidens** | |
|---|---|
| **Risiko for bias** | Vurdering af risiko for bias omhandler, i hvilken grad vi kan stole på resultaterne. Risikoen for bias i de enkelte studier vurderes ved hjælp af tjeklister, som beskrevet i afsnit 2.4. Herefter vurderes tilliden til evidensen for hvert effektmål på tværs af studier. |
| **Inkonsistens** | Inkonsistens omhandler ikke-forklarede forskelle i effektestimaterne på tværs af de inkluderede studier. Kriterier for vurdering af inkonsistens omfatter: lighed i effektstørrelse på tværs af studier, graden af overlap af konfidensintervaller samt statistiske test for heterogenitet |
| **Indirekte evidens** | Direkte evidens er evidens fra studier, som direkte sammenligner de interventioner, vi er interesserede i hos den patientgruppe, vi er interesserede i, og som har anvendt centrale patientrelaterede effektmål. Tiltroen til et givet resultat er størst, når evidensen er direkte. Afviger patientgruppen, interventionen eller effektmålene fra det, som vi er interesserede i at undersøge, kan det være nødvendigt at nedgradere grundet indirekte evidens. Ligeledes kan det være nødvendigt at nedgradere i de tilfælde, hvor der ikke findes direkte (head-to-head) sammenligninger mellem de interventioner, vi ønsker at vurdere. |
| **Unøjagtighed** | Unøjagtighed omhandler overordnet set, hvor præcise resultaterne er (hvor smalle de omtalte 95 % sikkerhedsgrænser er omkring metaanalysen), hvor meget data vi har, og hvor stor usikkerheden i resultaterne er. |
| **Publikationsbias** | Publikationsbias omhandler en vurdering af sandsynligheden for, at der foreligger upublicerede studier, som adskiller sig resultatmæssigt fra de publicerede studier i forhold til effekten af den intervention, vi ønsker at vide noget om. |

I det følgende gennemgås principperne for gradering af evidens for hvert af de fem kriterier, som kan føre til nedgradering af evidens.

### Risiko for bias

Efter at de inkluderede primære studier er blevet vurderet med hensyn til deres interne validitet ved hjælp af tjeklister, foretages en samlet vurdering af risikoen for bias på tværs af studier for hvert effektmål. Formålet med at vurdere risikoen for bias for hvert enkelt effektmål er at angive, i hvilken grad vi kan stole på resultaterne. Denne information kan blandt andet anvendes til at lægge større vægt på resultater fra studier med lavere risiko for bias [16]. For hvert effektmål foretages en vurdering af, hvorvidt kvaliteten af evidensen skal nedgraderes på baggrund af risikoen for bias [17].

I tabel 7 præsenteres strukturen ved gradering af evidens på baggrund af risiko for bias i randomiserede studier. Den anden kolonne i tabel 7 præsenterer tilgangen ved gradering af individuelle studier, mens de resterende kolonner refererer til den samlede evidens på tværs af de inkluderede studier (eng. body of evidence).

**Tabel 7. GRADE-struktur ved gradering af evidens på baggrund af risiko for bias [17]**

| Omfang af risiko for bias | Risiko for bias i det enkelte studie | Risiko for bias på tværs af studier | Fortolkning på tværs af studier |
|---|---|---|---|
| Ingen alvorlige begrænsninger: nedgrader ikke. | Lav risiko for bias i alle centrale domæner. | Hovedparten af information er fra studier med lav risiko for bias. | Høj tillid til evidens: Den sande effekt ligger tæt på effektestimatet. |
| Alvorlige begrænsninger: nedgrader et niveau (fx fra høj til moderat). | Alvorlige begrænsninger for et kriterium eller moderate begrænsninger for flere kriterier, tilstrækkeligt til at reducere tiltroen til effektestimatet. | Hovedparten af information er fra studier med moderat risiko for bias. | Tilliden til evidensen er reduceret fra høj til moderat: Den sande effekt ligger formodentlig tæt på effektestimatet, men der er nogen usikkerhed om resultaterne. |
| Meget alvorlige begrænsninger: nedgrader to niveauer (fx fra høj til lav). | Alvorlige begrænsninger i én eller flere kriterier, tilstrækkeligt til at reducere tiltroen til effektestimatet. | Hovedparten af information er fra studier med høj risiko for bias. | Tilliden til evidens reduceres fra høj til lav: Den sande effekt kan være betydeligt forskellig fra det analyserede effektestimatet. |

Overgangen fra vurdering af risiko for bias ved hjælp af en tjekliste i det enkelte studie til vurdering af nedgradering af evidens grundet risiko for bias på tværs af en række studier for et givet effektmål er en vanskelig opgave. I tabel 8 præsenteres principper, som kan anvendes til at guide processen [17].

**Tabel 8. Principper for nedgradering af evidens på baggrund af risiko for bias**

| Principper for nedgradering af evidens på baggrund af risiko for bias | |
|---|---|
| **1** | Den overordnede tillid til evidensen bør ikke fastsættes på baggrund af et gennemsnit af studiernes vurderinger. Der bør i stedet foretages en velovervejet vurdering af bidraget fra hvert studie med en generel anbefaling om at fokusere på studier med god intern validitet. |
| **2** | Ovenstående forudsætter vurdering af, i hvilket omfang hvert studie bidrager til det samlede effektestimat. Omfanget af bidraget afspejler som oftest studiets 'sample size' og for binære effektmål antallet af 'events'. Større studier med mange events bidrager mere end mindre studier med få events. |
| **3** | Vurdering om nedgradering bør være konservativ. Dvs. man skal være sikker på, at der er en betydelig risiko for bias hen over det meste af den tilgængelige evidens, før der nedgraderes for risiko for bias. |
| **4** | Risikoen for bias skal vurderes i kontekst med andre begrænsninger. Hvis fx bedømmerne er meget tæt på at nedgradere i to domæner, fx risiko for bias og unøjagtighed, anbefales det, at der som minimum nedgraderes i et af domænerne. |
| **5** | Det kan ikke undgås, at der vil være tilfælde, hvor bedømmere vil være på vippen i forhold til, om der skal nedgraderes eller ej. I disse tilfælde bør dette anerkendes, og overvejelserne samt begrundelse for den endelige vurdering gøres eksplicitte. |

For mere detaljeret information vedr. nedgradering på baggrund af risiko for bias henvises til følgende artikel af Balshem et al.: GRADE guidelines: 3. Rating the quality of evidence [17].

### Inkonsistens

Inkonsistens omhandler ikke-forklarede forskelle i effektestimaterne på tværs af de inkluderede studier. Kriterier for vurdering af inkonsistens omfatter: lighed i effektstørrelse på tværs af studier, graden af overlap af konfidensintervaller samt statistiske test for manglende homogenitet [29].

Betydelige forskelle i effekten af en given behandling mellem studier kan skyldes forskelle i: populationen (sygdomsgrad), behandling (dosis, co-interventioner, sammenligning), effekt (tid til followup) og metode (risiko for bias). Såfremt én af de første tre kategorier kan forklare forskellene i resultaterne, bør resultaterne præsenteres for hver subgruppe separat [29]. Såfremt forskellige estimater mellem studier overbevisende kan forklares ud fra forskelle i den metodiske tilgang, fokuseres på resultaterne fra studier med lavere risiko for bias. Ovenstående søgen efter forklaringer på forskelle i effekt foretages på baggrund af nogle få a priori-genererede hypoteser udarbejdet af arbejdsgruppen. Forklarede forskelle giver således ofte et argument for ikke at nedgradere tilliden til evidensen [29].

Såfremt der er betydelig ikke-forklaret variation i de inkluderede studiers resultater, vil det være hensigtsmæssigt at overveje at nedgradere kvaliteten af evidensen – særligt hvis nogle studier viser betydelig gavnlig effekt af et givet tiltag, mens andre viser ingen effekt eller ligefrem skadelig virkning [29]. I tabel 9 præsenteres kriterier, som bør medføre overvejelser omkring nedgradering på baggrund af inkonsistens.

**Tabel 9. Principper for nedgradering af evidens**

| | **Principper for nedgradering af evidens på baggrund af inkonsistens** |
|---|---|
| **1** | Punktestimaterne varierer betydeligt mellem studier |
| **2** | Der er minimal eller ingen overlap af konfidensintervaller (CI). |
| **3** | Den statistiske test for heterogenitet – som tester nulhypotesen om, at alle studier i metaanalysen har den samme underliggende effektstørrelse – har en lav p-værdi. |
| **4** | $I^2$ – som kvantificerer procentdelen af den totale variation mellem studierne som kan tillægges inkonsistens (heterogenitet) og ikke tilfældighed – er betydelig*. |

*Der findes ikke nogen fast definition for, hvad der skal til for at $I^2$ kan siges at være betydelig, men som tommelfingerregel betragtes en $I^2$ på > 50 % som værende betydelig. $I^2$ har som andre statistiske test for heterogenitet visse begrænsninger, og resultatet af testen skal derfor altid ses i kontekst med den subjektive vurdering af punktestimater og overlap af konfidensintervaller [29].

For mere detaljeret information vedr. nedgradering på baggrund af inkonsistens henvises til følgende artikel af Guatt GH et al.: GRADE guidelines: 7. Rating the quality of evidence-inconsistency [29].

### Indirekte evidens

Direkte evidens er evidens fra studier, som direkte sammenligner de interventioner, vi er interesserede i hos den patientgruppe, vi er interesserede i, og som har anvendt centrale patientrelaterede effektmål. Tiltroen til et givet resultat er størst, når evidensen er direkte. Afviger populationen (patientgruppen), interventionen eller effektmålene fra det, som vi er interesserede i at undersøge, kan det være nødvendigt at nedgradere grundet indirekte evidens. En fjerde – anderledes – form for indirekte evidens forekommer i tilfælde, hvor der ikke findes direkte (head-to-head) sammenligninger mellem de interventioner, vi ønsker at vurdere [13]. I tabel 10 gives en kort gennemgang af principperne for vurdering af indirekte evidens for hvert af de fire områder.

**Tabel 10. Principper for vurdering af indirekte evidens**

| **Principper for vurdering af indirekte evidens** | |
|---|---|
| **Population** | Studiepopulationen er defineret på forhånd, og det vil derfor kun være de studier, der opfylder inklusionskriterierne vedr. studiepopulationen, som inkluderes. Evidensen vedr. populationen (patientgruppen) vil derfor som hovedregel være direkte. Der kan dog være tilfælde, hvor det grundet manglende evidens vil være nødvendigt at inddrage evidens fra en patientgruppe, som adskiller sig i større eller mindre grad fra den gruppe, vi ønsker at vide noget om [13]. Det kunne fx være, at vi ønskede at undersøge effekten af et givet tiltag til patienter over 65 år, men at der kun eksisterer evidens for effekten til patienter i alderen 18-65 år. I dette tilfælde kan man overveje at nedgradere tilliden til evidensen. Generelt skal der kun nedgraderes på baggrund af populationsforskelle, såfremt der er overbevisende grundlag for at tro, at biologien hos den population, vi ønsker at sige noget om, er så forskellig fra testpopulationen, at effektstørrelsen vil være betydeligt anderledes. Som oftest vil dette ikke være tilfældet [13[. |
| **Intervention** | Ligesom ved studiepopulationen vil interventionen som hovedregel være direkte evidens, idet kun studier, som opfylder inklusionskriterierne i forhold til intervention, inkluderes i vurderingen. Igen vil der dog være tilfælde, hvor dette ikke er gældende, og hvor det derfor skal vurderes, om tilliden til evidensen skal nedgraderes [13]. |
| **Effektmål** | Centrale patientrelaterede effektmål bør være defineret på forhånd. De tilgængelige studier kan have undersøgt effekten af en given intervention på effektmål, som er relateret til – men anderledes – end de effektmål, som er vigtigst for patienterne. Disse afvigelser mellem de ønskede og de målte effektmål kan fx være relateret til tidshorisonten. Ofte vil man gerne undersøge effekten af en given intervention på lang sigt, fx efter 12 måneder. Såfremt der kun findes studier med opfølgning efter seks måneder, kan dette føre til nedgradering. En anden kilde til indirekte evidens relateret til effektmålet er anvendelse af surrogatmål som stedfortræder for det patientrelaterede effektmål, som vi ønsker at vide noget om. Generelt vil anvendelse af surrogatmål medføre nedgradering et eller to niveauer. Ved beslutning om, hvorvidt der skal nedgraderes et eller to niveauer, kan overvejelser om den formodede kausale sammenhæng mellem surrogatmålet og det patientrelaterede effektmål, vi ønsker at vide noget om, anvendes til at guide beslutningen. I yderst sjældne tilfælde hvor et surrogatmål gentagne gange er påvist at være tæt relateret til ændringer i et givet patientrelateret effektmål i RCT'er, kan det undlades at nedgradere evidensen på baggrund af indirekte evidens [13]. |

| | |
|---|---|
| **Indirekte sammenligninger** | Den sidste type af indirekte evidens forekommer, når der ikke findes direkte (head-to-head) sammenligninger mellem de to eller flere typer af tiltag, vi ønsker at vurdere – som hvis vi fx ønsker at sammenligne to lægemidler, A og B, og der ikke findes randomiserede studier, som sammenholder effekten af A med B, men kun studier som sammenholder A med placebo og B med placebo. Sådanne studier gør indirekte sammenligning af effekten af de to lægemidler mulig. Tilliden til evidensen er dog lavere end ved direkte sammenligninger [13]. Indirekte sammenligninger kan være meget vanskelige at håndtere, idet det kan være vanskeligt at vurdere, om forskellen i den relative risikoreduktion skyldes forskelle i effekten af de to lægemidler, eller om forskellen mere skal tillægges metodiske forskelle mellem studier, fx forskellige in- og eksklusionskriterier eller forskellige co-interventioner. Validiteten af de indirekte sammenligninger hviler på antagelsen om, at faktorer vedr. studiets design og metodiske kvalitet ikke er så forskellige, at det resulterer i forskellig effekt. Med andre ord – den angivelige forskel mellem studiernes resultater skyldes en sand forskel i effekten af de to interventioner. Eftersom der altid vil være nogen usikkerhed omkring denne antagelse, vil indirekte sammenligninger altid berettige nedgradering af evidensen. Hvorvidt, der skal nedgraderes et eller to niveauer, afhænger af i hvor høj grad, vi tror, at alternative faktorer (population, intervention, co-intervention, effektmål og metode) kan forklare eller sløre forskelle i effekten [13]. Vurderingen af tilliden til evidensen ved indirekte sammenligninger gøres endnu mere vanskelig ved anvendelse af statistiske tilgange til indirekte sammenligninger – herunder netværksmetaanalyser. |

Når alle fire former for indirekte evidens er gennemgået, foretages en vurdering. Hvis der er problemer med mere end én type af indirekte evidens, bør det overvejes at nedgradere to niveauer. Denne overvejelse er ikke en simpel additiv proces, men mere en vurdering af, hvor meget nedgradering der er nødvendig. Generelt bør evidens fra surrogatmål føre til nedgradering, hvorimod andre former for indirekte evidens vil forudsætte en mere velovervejet vurdering [13].

For mere detaljeret information vedr. nedgradering på baggrund af indirekte evidens henvises til følgende artikel af Guatt GH et al.: GRADE guidelines: 8. Rating the quality of evidence-indirectness [13].

**Unøjagtighed**

Unøjagtighed omhandler overordnet set, hvor præcise resultaterne er, hvor meget data vi har, og hvor stor usikkerheden i resultaterne er [30].

Det primære kriterium for vurdering af unøjagtighed i GRADE er for hvert effektmål at fokusere på 95 %-konfidensintervallet (CI) omkring forskellen i effektestimatet mellem intervention og sammenligningsalternativet [30]. 95 %-konfidensintervallet udtrykker, at den sande værdi, der ønskes estimeret med 95 % sandsynlighed, ligger i CI. Når kvaliteten af evidensen skal bestemmes, skal det vurderes, hvorvidt CI omkring et givet effektestimat er tilstrækkeligt smalt.

Hvis CI vurderes ikke at være tilstrækkelig smalt, nedgraderes evidensen et niveau (fx fra høj til moderat). Hvis CI er meget brede, kan der nedgraderes to niveauer [30].

Vurdering af, hvorvidt konfidensintervallet er tilstrækkeligt smalt, foregår i første omgang ved at se på, hvorvidt den øvre (eller nedre) grænse af konfidensintervallet overskrider den kliniske tærskelværdi for anbefaling af den givne intervention. Såfremt den nedre eller øvre grænse af vores konfidensinterval ligger over den kliniske tærskelværdi for anbefaling af den givne intervention, bør kvaliteten af evidensen nedgraderes. Såfremt den øvre (eller nedre) ende af konfidensintervallet ikke krydser tærskelværdien for, hvorvidt man vil anbefale den givne intervention, foretages der en vurdering af informationsstørrelsen, som ligger til grund for estimatet og det tilhørende konfidensinterval. Dette skyldes blandt andet, at et højt effektestimat med et smalt konfidensinterval i lille studie med relativt få 'events' kan tillægges tilfældigheder. Fx vil studier, som stoppes tidligt grundet imponerende positive resultater, ofte overestimere behandlingens effekt. Derfor skal der ved vurdering af unøjagtighed ud over CI også indgå en vurdering af informationsstørrelsen, som ligger til grund for CI. Operationaliseringen af dette foregår ved at beregne den optimale informationsstørrelse. Såfremt det totale antal af patienter, som indgår i litteraturgennemgangen, er mindre end det antal patienter, som ved beregning af konventionel 'sample size', er nødvendigt, for at et enkelt studie har adækvat statistisk power, nedgraderes kvaliteten af evidensen [30]. Der findes adskillige onlineredskaber til beregning af 'sample size'. I tilfælde hvor 'sample size' er meget stor (ved få events), skal der ikke nedgraderes, til trods for at den optimale 'sample size' ikke er opnået.

For mere detaljeret information vedr. nedgradering på baggrund af unøjagtighed henvises til følgende artikel af Guatt GH et al.: GRADE guidelines: 6. Rating the quality of evidence-imprecision [30].

**Publikationsbias**

Publikationsbias omhandler en vurdering af sandsynligheden for, at der foreligger upublicerede studier, som adskiller sig resultatmæssigt fra de publicerede studier i forhold til effekten af den intervention, vi ønsker at vide noget om [31]. Denne disciplin er omgivet af en vis mystik, da det kræver en del detektivarbejde at sandsynliggøre 'publication bias'. Der er empirisk dokumentation for, at studier med statistisk signifikante resultater i højere grad bliver publiceret end studier uden statistisk signifikante resultater. Derudover er der en tendens til, at der går længere tid fra at studier med negative resultater eller 'null findings' bliver publiceret sammenholdt med studier med positive fund. Dette kan medvirke til, at resultatet af en given intervention overestimeres – særligt når litteraturgennemgangen foregår i en tidlig fase, hvor der kun findes få studier med lille 'sample size', som kan dokumentere effekten af den givne interven-

tion (31). Derudover bør der være særlig opmærksomhed på risiko for publikationsbias, såfremt de få tilgængelige studier, der foreligger, er sponsoreret af industrien [31].

Det er meget vanskeligt at kunne udelukke publikationsbias og næsten ligeså vanskeligt at fastsætte en tærskelværdi for, hvornår der skal nedgraderes for publikationsbias. Der findes flere metoder, som kan anvendes til at vurdere sandsynligheden for publikationsbias. Den mest populære af disse metoder er de såkaldte 'funnel Plots'. Anvendelse og fortolkning af 'funnel plots' har dog betydelige begrænsninger, idet den statistiske validitet af testen betvivles af en række eksperter på området [32]. Grundet den store usikkerhed ved gradering inden for dette kriterium anvendes termerne 'ikke detekteret' og 'stærk mistanke'. I erkendelse af denne usikkerhed bør mistanke om publikationsbias kun medføre nedgradering et niveau (frem for to). Ikke desto mindre er publikationsbias sandsynligvis hyppigt forekommende, særligt på områder hvor størstedelen af forskningen kommer fra studier sponsoreret af industrien [31].

For mere detaljeret information vedr. nedgradering på baggrund af publikationsbias henvises til følgende artikel af Guatt GH et al.: GRADE guidelines: 5. Rating the quality of evidence-publication bias [31].

#### 2.6.1.2 Opgradering af evidens

Observationelle studier klassificeres som udgangspunkt til lav tillid til evidensen. Der kan dog være tilfælde, hvor vi har høj tiltro til estimater fra kohortestudier på grund af særlige karakteristika. Derfor har man i GRADE beskrevet tre kriterier for opgradering af kvaliteten af evidens i kohortestudier. De tre kriterier, som kan føre til opgradering, er præsenteret i tabel 11.

**Tabel 11. Faktorer som kan øge kvaliteten af evidens**

| **Faktorer som kan øge kvaliteten af evidens** | |
|---|---|
| **Effektstørrelse** | Stor: direkte evidens, RR = 2-5 eller tilsvarende RR 0,5-0,2 uden plausibel konfounding.<br>Meget stor: RR >5 eller tilsvarende RR<0,2 og ingen alvorlige problemer med risiko for bias eller unøjagtighed. |
| **Dosis-respons-sammenhæng** | Der er tegn på dosis-respons-sammenhæng. |
| **Konfounding** | Al plausibel residualkonfounding eller bias ville reducere den påviste effekt eller medføre effekt, hvis resultatet viser 'ingen effekt'. |

Kriteriet 'effektstørrelse' kan medføre opgradering af tilliden til evidensen med et niveau ved stor effektstørrelse, eller to niveauer ved meget stor effektstørrelse (fx fra lav til moderat eller fra lav til høj). De to sidste kriterier kan medføre opgradering med et niveau [19].

### 2.6.2 Vurdering af den samlede kvalitet af evidens for hvert enkelt effektmål

Når de fem kriterier for nedgradering og de tre kriterier for opgradering af kvaliteten af evidens er gennemgået, skal der foretages en samlet vurdering af kvaliteten af evidens for hvert af de kritiske eller vigtige effektmål. Den samlede kvalitet af evidensen for hvert effektmål kan graderes i fire niveauer: Høj, moderat, lav eller meget lav. Vurderingen af den samlede kvalitet af evidensen for det enkelte effektmål beror på en overordnet vurdering af tiltroen til evidensen. I mange tilfælde vil indplacering i en given kategori af evidens, baseret på en simpel optælling af nedgraderinger, ikke give et retvisende billede af vores overordnede tiltro til kvaliteten af evidens [33].

I tabel 12 ses definitionerne for de fire kategorier af evidens i GRADE.

**Tabel 12. Definition af de fire evidenskategorier i GRADE**

| Tillid | Definition |
|---|---|
| Høj<br>++++ | Vi er meget sikre på, at den sande effekt ligger tæt på vores effektestimat. |
| Moderat<br>+++ | Vi er moderat sikre på effektestimatet: Den sande effekt er sandsynligvis tæt på effektestimatet, men der er en risiko for, at den sande effekt i virkeligheden er anderledes end effektestimatet. |
| Lav<br>++ | Vores tillid til effektestimatet er begrænset: Den sande effekt kan være væsentlig anderledes end effektestimatet. |
| Meget lav<br>+ | Vi har meget lav tillid til effektestimatet: Den sande effekt kan meget vel tænkes at være væsentlig anderledes end effektestimatet. |

### 2.6.3 Evidensprofiler

Resultaterne af GRADE-gennemgangen præsenteres i en såkaldt evidensprofil. Evidensprofilen er en tabel, som illustrerer vurderingen for hvert af de kritiske eller vigtige effektmål samt den samlede vurdering af evidensen. Derudover indeholder evidensprofilerne information om antal patienter og events samt effektforhold fra de inkluderede studier. Evidensprofilerne kan udarbejdes ved hjælp af GRADEpro-software. I figur 9 ses et eksempel på en evidensprofil.

**Figur 9. Eksempel på en evidensprofil**

| Kvalitetsvurdering | | | | | | | Antal patienter og events inkluderet i analyserne | | Effekt | | Over-ordnet kvalitet |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Effektmål (antal studier) | Studie-design | Risiko for bias | Inkonsi-stens | Indirekte evidens | Unøjagtig-hed | Publika-tionsbias | Robot-assisteret | Laparo-skopi | Relativ forskel (95 % CI) | Gennemsnitlig forskel (CI) | |
| **Robotassisteret kirurgi over for laparoskopi** | | | | | | | | | | | |
| Operations-tid (5) | RK | Alvorlig | Alvorlig | Ikke alvorlig | Ikke alvorlig | Ikke observeret | 744 | 788 | - | MD: 5,9 min. længere ved robotkirurgi (CI:-32,10; 43,91) | ⊕OOO Meget lav |
| Operations-tid (2) | RCT | Ikke alvorlig | Ikke alvorlig | Ikke alvorlig | Alvorlig | Ikke observeret | 73 | 73 | - | MD: 44,25 min. længere ved robotkirurgi (CI: 5,21; 83,30) | ⊕⊕⊕O Moderat |
| Indlæggel-sestid (1) | RK | Alvorlig | Ikke alvorlig | Ikke alvorlig | Ikke alvorlig | Ikke observeret | 237 | 265 | - | MD: 0,2 dage kortere ved robotkirurgi (-0,29; 0,11) | ⊕OOO Meget lav |

RK: retrospektivt kohortestudie

### 2.6.4 Vurdering af den samlede kvalitet af evidens

Når den samlede tillid til evidensen er vurderet for hvert effektmål, foretages en vurdering af den samlede tiltro til evidensen på tværs af effektmål. I den samlede vurdering tages der udgangspunkt i de effektmål, som er vurderet som værende kritiske. Det samlede evidensniveau kan som udgangspunkt ikke være højere end det lavest vurderede kritiske effektmål [33].

For mere detaljeret information vedr. vurdering af kvaliteten af evidens på tværs af effektmål henvises til følgende artikel af Guatt GH et al.: GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes (33).

### 2.6.5 Hvad gør man, hvis evidensgrundlaget kun består af et enkelt RCT-studie?

Når nye teknologier eller tiltag vurderes, vil der ofte kun være ganske få eller måske kun et enkelt studie til rådighed. Set i lyset af, at en given positiv effekt beskrevet i ét studie i mange tilfælde ikke kan genfindes i efterfølgende studier, vil man (med rette) som bedømmer kunne

føle sig utryg ved at betegne ét RCT-studie som 'høj tillid til evidensen'. På den anden side vil det ikke være hensigtsmæssigt at nedgradere kvaliteten af evidens automatisk i de tilfælde, hvor der kun findes et studie. Et enkelt stort veltilrettelagt og veludført multicenter-RCT kan levere høj kvalitet af evidens (typisk vil den manglende evidens afsløres ved brede 95 % sikkerhedsgrænser). I tilfælde, hvor der kun findes et studie, anbefales særlig grundig gransk-ning af alle relevante domæner (risiko for bias, indirekte evidens, unøjagtighed og publika-tionsbias) – da inkonsistens i sagens natur ikke kan vurderes [17].

# 3 Peer-review og publicering

En ekstern vurdering, også kaldet fagfællevurdering eller peer-review, er en procedure som typisk anvendes i forbindelse med publicering af videnskabelige artikler med henblik på at sik-re eller forbedre kvaliteten af forskningsproduktet. I DEFACTUM ønsker vi i størst mulig grad at indhente eksterne vurderinger af vores litteraturgennemgange – ikke mindst for at sikre at beskrivelserne af de kliniske aspekter bliver så god som mulig. Vi anvender også intern peer-review, særligt med henblik på at sikre den metodiske kvalitet af produktet.

# 4 Publicering

Alle produkter publiceres på DEFACTUMs hjemmeside. Videnskabelige artikler publiceres i relevante, anerkendte videnskabelige tidsskrifter og EUnetHTA-rapporter publiceres på EUnetHTA's hjemmeside.

# 5 Referencer

(1) Kristensen FB, Sigmund H (red.). Metodehåndbog for Medicinsk Teknologivurdering. København: Sundhedsstyrelsen, 2007.

(2) Schunemann HJ, Tugwell P, Reeves BC et al. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. Research synthesis methods. Res Synth Methods 2013;4(3):287-9.

(3) Guyatt GH, Oxman AD, Kunz RF et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011;64(4):395-400.

(4) EUnetHTA Joint Action WP5 – Relative Effectiveness Assessment (REA) of Pharmaceuticals. Model for Rapid Relative Effectiveness Assessment of Pharmaceuticals– V3.0. EUnetHTA, March 2013

(5) Sundhedsstyrelsen. METODEHÅNDBOGEN. Model for udarbejdelse af nationale kliniske retningslinjer. København: Sundhedsstyrelsen, 2018.

(6) EUnetHTA Guidelines. COMPARATORS & COMPARISONS. Criteria for the choice of the most appropriate comparator(s). Summary of current policies and best practice recommendations. EUnetHTA, 2013.

(7) EUnetHTA Guidelines. Endpoints used for relative effectiveness assessment of pharmaceuticals: CLINICAL ENDPOINTS. EUnetHTA, 2013.

(8) Christensen R, Maxwell LJ, Juni P et al. Consensus on the Need for a Hierarchical List of Patient-reported Pain Outcomes for Metaanalyses of Knee Osteoarthritis Trials: An OMERACT Objective. J Rheumatol 2015;42(10):1971-5.

(9) Ioannidis JP, Evans SJ, Gøtzsche PC et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med 2004;141(10):781-8.

(10) Zorzela L, Loke YK, Ioannidis JP et al. PRISMA harms checklist: improving harms reporting in systematic reviews. BMJ 2016;352:i157.

(11) EUnetHTA Guideline. Endpoints used for relative effectiveness assessment: Health-related quality of life and utility measures. EUnetHTA, 2015.

(12) Bucher HC, Guyatt GH, Cook DJ et al. Users' guides to the medical literature: XIX. Applying clinical trial results. A. How to use an article measuring the effect of an intervention on surrogate end points. Evidence-Based Medicine Working Group. JAMA 1999;282(8):771-8.

(13) Guyatt GH, Oxman AD, Kunz R et al. GRADE guidelines: 8. Rating the quality of evidence-indirectness. J Clin Epidemiol 2011;64(12):1303-10.

(14) Institute for Quality and Efficiency in Health Care (IQWIG). General methods. Køln, 2015.

(15) Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.handbook.cochrane.org.

(16) Nasjonalt kunnskapssenter for helsetjenesten. Slik oppsummerer vi forskning. Håndbok for Nasjonalt kunnskapssenter for helsetjenesten. 4. reviderte utg. Oslo: Nasjonalt kunnskapssenter for helsetjenesten, 2015.

(17) Balshem H, Helfand M, Schunemann HJ et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;64(4):401-6.

(18) Murad MH, Asi N, Alsawas M et al. New evidence pyramid. Evid Based Med 2016;21(4):125-7.

(19) Guyatt GH, Oxman AD, Sultan S et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 2011;64(12):1311-6.

(20) Shamseer L, Moher D, Clarke M et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation BMJ 2015;350:g7647.

(21) Liberati A, Altman DG, Tetzlaff J et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009;339:b2700.

(22) Higgins JP, Altman DG, Gøtzsche PC et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.

(23) Sterne JA, Hernan MA, Reeves BC et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919.

(24) Guyatt GH, Oxman AD, Santesso NF et al. GRADE guidelines: 12. Preparing summary of findings tables-binary outcomes. J Clin Epidemiol 2013;66(2):158-72.

(25) Guyatt GH, Thorlund K, Oxman AD et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. J Clin Epidemiol 2013;66(2):173-83.

(26) Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med 2002;21(11):1575-600.

(27) Guyatt GH, Oxman AD, Vist GF et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). J Clin Epidemiol 2011;64(4):407-15.

(28) Guyatt GH, Oxman AD, Akl EA et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64(4):383-94.

(29) Guyatt GH, Oxman AD, Kunz RF et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol 2011;64(12):1294-302.

(30) Guyatt GH, Oxman AD, Kunz RF et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. J Clin Epidemiol 2011;64(12):1283-93.

(31) Guyatt GH, Oxman AD, Montori V et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. J Clin Epidemiol 2011;64(12):1277-82.

(32) Sterne JA, Sutton AJ, Ioannidis JP et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011;343:d4002.

(33) Guyatt G, Oxman AD, Sultan SF et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. J Clin Epidemiol 2013;66(2):151-7.