Masteroppgave

Fysioterapi for muskelskjeletthelse,

master i helsevitenskap

Mai 2024

FREMTIDENS FORSKNING?

Bruk av maskinlæringsverktøy for studieseleksjon til

systematiske oversiktsartikler

Kandidatnavn: Renate Haugland Solheim
Emnekode: MAVITD5900

Antall ord: 18 296

Fakultet for helsevitenskap
OSLO METROPOLITAN UNIVERSITY
STORBYUNIVERSITETET

Forord

Det er med både stolthet og lettelse at jeg presenterer denne masteroppgaven. De siste tre årene har bydd på oppturer og nedturer, både personlig og faglig. Veien hit har gitt meg flere utfordringer og har ikke alltid vært enkel, noe som gjør at jeg i dag er ekstra takknemlig og stolt over å ha nådd målstreken.

Først og fremst vil jeg takke CIM ved OsloMet for et spennende masterprosjekt, og min veileder Bjørnar Berg for all god hjelp på veien. Jeg kastet meg inn i et prosjekt jeg hadde lite forkunnskaper om, og uten din tålmodighet og gode tilbakemeldinger ville oppgaven ikke blitt den samme.

Et stort takk går også til mine medstudenter og kolleger på de to arbeidsplassene jeg har hatt i denne perioden. Jeg har satt stor pris på all tilrettelegging og hjelp jeg har fått med mine arbeidsoppgaver for at jeg skulle kunne fullføre denne mastergraden.

Til mamma, pappa og Henning: Takk for at dere er mine viktigste støttespillere, og for at dere alltid heier på meg uansett hva jeg gjør.

Til slutt vil jeg også takke både nye og gamle venner, som har betydd så utrolig mye for meg disse årene. Dere har bidratt med mange nødvendige avbrekk fra skrivingen.

Jeg håper denne oppgaven vil bidra til å kaste lys over et spennende emne, og kanskje inspirere andre til å utforske det videre.

Takk for at du tar deg tid til å lese!

Renate Haugland Solheim

Oslo, 14. mai 2024

Innholdsfortegnelse

Sammendrag

Bakgrunn: For å sikre tilgang til høyest mulig grad av evidens i klinisk praksis har det blitt løftet frem et behov for hurtigere produksjon av systematiske oversiktsartikler. Kunstig intelligens kan potensielt benyttes for å automatisere deler av prosessen. Dette er særlig aktuelt ved studieseleksjon, hvor arbeidsbelastningen kan reduseres med 30-70%. Formålet med oppgaven var derfor å vurdere prestasjonen til maskinlæringsverktøyet i et nettbasert verktøy (Rayyan) for automatisering av studieseleksjonen til en systematisk oversiktsartikkel om prognostiske modeller for degenerativ ryggkirurgi.

Metode: Evnen maskinlæringsverktøyet har til å identifisere relevante artikler ble sammenlignet med menneskelige vurderinger. De 7994 aktuelle artiklene ble manuelt sortert til kategoriene inkludert eller ekskludert i 20% nivåer. Etter hvert nivå ble maskinlæringsverktøyet brukt til å rangere de resterende artiklene etter relevans (fra 0.5 til 4.5 stjerner). To ulike grenseverdier for eksklusjon av artikler ble brukt: <2.5 og ≤2.5 stjerner. Utfallsmålene som ble brukt for å vurdere prestasjonen var sensitivitet, spesifisitet, positiv prediktiv verdi og negativ prediktiv verdi.

Resultat: Med en grenseverdi på <2.5 stjerner ble best prestasjon oppnådd etter manuell sortering av 60% av studiene, med en sensitivitet på 100% og spesifisitet på 68%. Allerede etter 20% manuell sortering var sensitiviteten over 96%, men spesifisiteten 38%. Tilsvarende var negativ prediktiv verdi høy og positiv prediktiv verdi lav ved alle nivåer. Dersom også artiklene med 2.5 stjerner ble ekskludert oppnådde maskinlæringsverktøyet nær perfekt spesifisitet ved alle nivåer (≥99.8%), men maksimalt 54.6% sensitivitet (etter sortering av 60%).

Konklusjon: Maskinlæringsverktøyet i Rayyan presterte godt nok til å kunne automatisk ekskludere en andel irrelevante artikler med en grenseverdi på <2.5 stjerner, og kan dermed potensielt redusere arbeidsbelastningen ved studieseleksjon til en systematisk oversiktsartikkel på prognostiske modeller for degenerativ ryggkirurgi. Menneskelige vurderinger er likevel fremdeles i stor grad nødvendig og videre utvikling av verktøyet er essensielt før full automatisering av oppgaven.

Stikkord: Automatisering, studieseleksjon, systematisk oversiktsartikkel, maskinlæring, Rayyan

Abstract

Background: To ensure access to the highest possible level of evidence in clinical practice, there is a need for faster production of systematic reviews. Artificial intelligence can potentially be used to automate parts of the process. This is particularly relevant in study selection, where workload can be reduced by 30-70%. The aim of this study was therefore to assess the performance of the machine learning tool in a web-based tool (Rayyan) for automating the study selection process for a systematic review on prognostic models for degenerative spine surgery.

Method: The ability of the machine learning tool to identify relevant articles was compared with human assessments. The 7994 relevant articles were manually categorized as included or excluded in 20% increments. After each increment the machine learning tool were used to rank the remaining articles by relevance (from 0.5 to 4.5 stars). Two different thresholds for exclusion were used: <2.5 and ≤2.5 stars. The outcome measures used to assess performance were sensitivity, specificity, positive predictive value and negative predictive value.

Results: With a threshold of <2.5 stars, best performance was achieved after manually sorting of 60% of the studies, with a sensitivity of 100% and specificity of 68%. Already after sorting 20%, sensitivity was over 96%, but specificity was 38%. Similarly, the negative predictive value was high and the positive predictive value was low at all levels. If articles with 2.5 stars were also excluded, the machine learning tool achieved near-perfect specificity at all levels (≥99.8%), but sensitivity reached a maximum of 54.6% (after sorting 60%).

Conclusion: The machine learning tool in Rayyan performed well enough to automatically exclude a portion of irrelevant articles with a threshold of <2.5 stars, and thus potentially reduce the workload in study selection for a systematic review on prognostic models for degenerative spine surgery. However human assessments are still largely necessary, and further development of the tool is essential before full automation of the task.

Keywords: Automation, study selection, systematic review article, machine learning, Rayyan

Forkortelser

KI – Kunstig intelligens

PPV – Positiv prediktiv verdi

NPV – Negativ prediktiv verdi

Begrepsavklaringer

Kunnskapsbasert praksis: Kunnskapsbasert praksis innebærer å basere fagutøvelsen på relevant evidensbasert kunnskap av god kvalitet, erfaringsbasert kunnskap og pasientenes preferanser.

Systematisk oversiktsartikkel: Litteraturstudier som gir et samlet bilde av kunnskapsgrunnlaget innen et avgrenset forskningsspørsmål ved å oppsummere allerede publisert forskning.

Kunstig intelligens (KI): Kunstig intelligente systemer utfører handlinger, fysisk eller digitalt, basert på tolkning og behandling av strukturerte eller ustrukturerte data, i den hensikt å oppnå et gitt mål. Enkelte KI-systemer kan også tilpasse seg gjennom å analysere og ta hensyn til hvordan tidligere handlinger har påvirket omgivelsene. Maskinlæring og naturlig språkbehandling er eksempler på former for KI.

Sensitivitet: Andelen artikler som ble vurdert til å være relevante for inklusjon av Rayyan blant studiene som faktisk ble inkludert ved menneskelig vurdering.

Spesifisitet: Andelen artikler som ble vurdert av Rayyan til å ikke være aktuell for inklusjon blant artiklene som ble ekskludert ved menneskelig vurdering.

PPV: Sannsynligheten for at artikkelen skal inkluderes når den vurderes som aktuell for inklusjon av Rayyan.

NPV: Sannsynligheten for at artikkelen skal ekskluderes når den ikke vurderes som aktuell for inklusjon av Rayyan.

1. Introduksjon

Kunnskapsbasert praksis er et viktig prinsipp for å bedre pasientbehandlingen og styrke beslutningsgrunnlaget innen fysioterapi. Kunnskapsbasert praksis innebærer å ta faglige avgjørelser basert på en kombinasjon av systematisk innhentet evidensbasert kunnskap, erfaringsbasert kunnskap og pasientens ønsker og behov (helsebiblioteket, 2021). Det har imidlertid lenge vært utfordringer knyttet til å overføre evidensbasert kunnskap til klinisk praksis, og studier har estimert at det tar 17 år før denne kunnskapen benyttes i klinikkene (Balas & Boren, 2000). Med omtrent 2 millioner årlige publiseringer om medisinske spørsmål er det behov for målrettede tiltak for å holde helsepersonell oppdatert på nyeste forskning (Balas & Boren, 2000).

Behovet for faglig oppdatering i en travel hverdag gjør lesing av oppsummert kunnskap mer hensiktsmessig enn enkeltstudier. Det å søke etter oppsummert forskning som systematiske oversiktsartikler før man går videre til enkeltstudier er et viktig prinsipp i kunnskapsbasert praksis (Jamtvedt et al., 2003). Utarbeidingen av systematiske oversiktsartikler er svært tidkrevende med dagens metoder, og det kan ta opptil 7 år før enkeltstudier blir implementert i en systematisk oversikt (Elliott et al., 2014; Jonnalagadda et al., 2015; Van De Schoot et al., 2021). For å sikre at klinikere har tilgang til best mulig evidens er det viktig at forskning syntetiseres uten unødvendig forsinkelser. Det har derfor blitt løftet frem et behov for hurtigere utvikling av systematiske oversiktsartikler (Schünemann & Moja, 2015). Det er nødvendig å ta i bruk nye og innovative metoder for å utvikle troverdig og betydningsfull forskningsbasert kunnskap. Interessen for å benytte kunstig intelligens (KI) har vært økende de siste årene, og teknologien viser lovende muligheter (Balas & Boren, 2000; Regjeringen, 2020).

De potensielle mulighetene som følger med den teknologiske utviklingen av KI har fått stor oppmerksomhet i nyhetsbildet etter lanseringen av ChatGPT i november 2022 (OpenAI, 2022). ChatGPT er basert på naturlig språkbehandling, en teknologi som også potensielt kan benyttes som et verktøy ved utarbeidelsen av systematiske oversiktsartikler. Etter testing av programvaren til dette formålet konkluderte Qureshi et al (2023) med at teknologien er lovende, men at det er stort behov for videre utvikling før slike programvarer kan tas i bruk med sikkerhet.

KI kan likevel allerede benyttes under utarbeidelsen av systematiske oversiktsartikler for å gjennomføre litteratursøk, vurdere sammendrag og ekstrahere resultater (Cierco Jimenez et al., 2022; Marshall & Wallace, 2019). Med dagens teknologi er det særlig relevant å anvende teknologien for å vurdere sammendrag fra systematiske søk for å identifisere artiklene som skal inkluderes i den systematiske oversikten (Cierco Jimenez et al., 2022; Marshall & Wallace, 2019). Det er likevel viktig å reflektere over om maskiner kan gjennomføre disse oppgavene uten å begå feil, om det er forsvarlig å overlate oppgaver til programvarer og hvem som er ansvarlig om det blir gjort feil.

Digitale verktøy som Covidence, Rayyan, Abstrackr og EPPI Reviewer er eksempler på programvarer som allerede relevante å bruke til studieseleksjonsprosessen (Cierco Jimenez et al., 2022; FHI, 2022). Flere av disse verktøyene har vist lovende prestasjoner i tidligere studier og kan redusere arbeidsbelastningen og tidsbruken for forskere som utarbeider systematiske oversiktsartikler (Chai et al., 2021; Tsou et al., 2020; Valizadeh et al., 2022). Rayyan er en av de mest populære programvarene for å organisere og gjennomgå studier i en systematisk oversikt da det er et gratis, nettbasert verktøy som ikke krever kunnskap om dataprogrammering. Det er også en av de oftest siterte programvarene (Cierco Jimenez et al., 2022). Tidligere studier poengterer likevel at det er behov for mer forskning for å kartlegge risikoer og fordeler ved å benytte maskinlæring, brukervennlighet, brukertilfredshet, kostnadseffektivitet og potensiell reduksjon i arbeidsbelastning (Tsou et al., 2020; Valizadeh et al., 2022). Den viktigste forutsetningen for at forskere skal kunne ta i bruk maskinlæringsverktøy ved studieseleksjon i systematiske oversikter er likevel at disse har god evne til å skille mellom artikler som skal inkluderes og ekskluderes. Det er derfor dette som utforskes i denne oppgaven.

2. Formål

Formålet med denne oppgaven er å vurdere prestasjonen til et maskinlæringsverktøy for automatisering av studieseleksjonen til en systematisk oversiktsartikkel. Dette er viktig på grunn av utfordringene knyttet til treffsikkerheten til søkestrategier i systematiske oversiktsartikler, hvor søkeresultatet ofte inkluderer mange irrelevante treff. Bruk av et maskinlæringsverktøy for automatisk studieseleksjon kan potensielt føre til en betydelig reduksjon i arbeidsbelastningen. Imidlertid forutsetter det at maskinlæringsverktøyet har god evne til å skille mellom relevante og irrelevante studier. For å undersøke dette vil litteratursøket fra en pågående systematisk oversiktsartikkel om prognostiske modeller for utfall etter degenerativ ryggkirurgi benyttes. Resultatene fra manuell studieseleksjon vil være sammenligningsgrunnlag for maskinlæringsverktøyet. Dette leder til følgende forskningsspørsmål: Hvordan er prestasjonen til det implementerte maskinlæringsverktøyet i Rayyan for å selektere studier for inklusjon og eksklusjon fra en søkestrategi på prognostiske modeller for utfall etter degenerativ ryggkirurgi?

Spesifikt vil prestasjonen til maskinlæringsverktøyet evalueres med sensitivitet, spesifisitet, positiv prediktiv verdi (PPV) og negativ prediktiv verdi (NPV).
Basert på tidligere forskning er min hypotese at maskinlæringsverktøyet vil oppnå høy sensitivitet, men lav spesifisitet (Olofsson et al., 2017; Valizadeh et al., 2022). Det vil si at det kan være et pålitelig verktøy for å ekskludere irrelevante studier, men at programvaren er mindre pålitelig når det gjelder å identifisere studiene som skal inkluderes.

3. Bakgrunn for valg av tema og teori

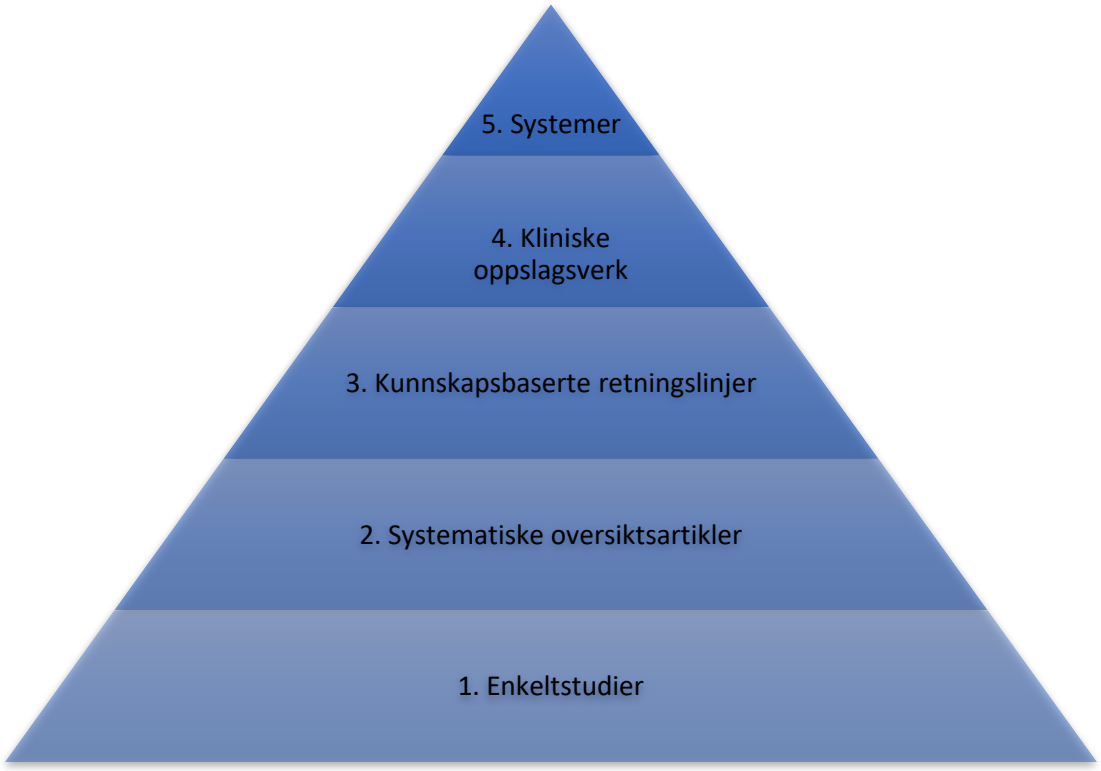
3.1. Kunnskapsbasert praksis

Det har aldri blitt publisert mer forskning enn i de siste tiårene, og fysioterapeuter opplever økende krav til å kunne håndtere den store mengden informasjon (Jamtvedt et al., 2003). Det ble i 2010 publisert 11 systematiske oversiktsartikler og 75 enkeltstudier daglig (Bastian et al.) Omfanget av denne utviklingen ser ut til å fortsette å øke, og årlig publiseres det flere millioner artikler innen det biomedisinske forskningsfeltet (Bastian et al., 2010; Greenhalgh, 2019). Bare i databasen PubMed finnes det over 37 millioner artikler og sammendrag (PubMed, 2024). Dette er uhåndterlige mengder informasjon for klinikere som ønsker å holde seg oppdatert på evidensbasert kunnskap.

Helsepersonelloven (1999, § 4) forplikter likevel fysioterapeuter til forsvarlig yrkesutøvelse. Som beskrevet i innledningen er det da viktig å følge prinsippene for kunnskapsbasert praksis, og benytte oppdatert evidensbasert kunnskap ved valg av behandlingsmetoder, for å sikre at pasientene tilbys helsetjenester av god kvalitet. Kunnskapsbasert fysioterapi innebærer å basere fagutøvelsen på relevant evidensbasert kunnskap av god kvalitet, erfaringsbasert kunnskap og pasientenes preferanser (Jamtvedt et al., 2003). Kunnskapsbasert praksis omfatter dermed mer enn bare forskning, men det presiseres at forskning bør være en sentral kilde til kunnskap i fagutøvelsen (Jamtvedt et al., 2003). Å holde seg faglig oppdatert kan være krevende av flere grunner, som for eksempel mangel på tid, fasiliteter, søkeferdigheter, motivasjon, mangel på kjennskap til gode informasjonskilder og uklare spørsmål (Greenhalgh, 2019; Jamtvedt et al., 2003). Det finnes flere nyttige verktøy for å overkomme disse barrierene og gjøre prosessen enklere.

Et eksempel på et slikt verktøy er kunnskapspyramiden (Figur 1), som illustrerer hvilke kunnskapskilder som er mest pålitelige og hvilken type kunnskapskilder det finnes mest av. Pyramiden kan dermed også illustrere hvor man bør begynne søket etter ny kunnskap. Når man som kliniker skal søke etter forskning ønsker man å finne pålitelige svar så raskt som mulig. For å sikre dette er det viktig å både vite hvor man skal begynne å lete og å kunne kritisk vurdere forskningen man finner. Når man benytter kunnskapspyramiden ønsker man å starte søket så høyt oppe i pyramiden som mulig. Dersom svaret man søker ikke finnes i retningslinjer eller kliniske oppslagsverk må man lete etter svar i studier. Som vist i pyramiden er systematiske

oversiktsartikler en bedre kunnskapskilde enn enkeltstudier (Helsebiblioteket, 2017). Disse artiklene danner også grunnlaget for videre utvikling av kunnskapsbaserte retningslinjer og kliniske oppslagsverk, som er trinn 3. og 4. i pyramiden (Akl et al., 2017). For å imøtekomme behovet til pasientene, klinikerne og politikere må andelen unødvendige enkeltstudier begrenses og utarbeiding av oppsummerende forskning som systematiske oversiktsartikler prioriteres (Bastian et al., 2010).



Figur 1: Illustrasjon av kunnskapspyramiden, modifisert fra (Helsebiblioteket, 2017)

3.1.1. Systematiske oversiktsartikler

Systematiske oversiktsartikler er litteraturstudier som gir et samlet bilde av kunnskapsgrunnlaget innen et avgrenset forskningsspørsmål ved å oppsummere allerede publisert forskning (Jamtvedt et al., 2003). For å bedre pasientbehandlingen i helsevesenet er det viktig å utforske metoder som kan bidra til å tette kunnskapshull og fremskynde implementeringsprosessen. De første systematiske oversiktene og metaanalysene i det medisinske forskningsfeltet ble utarbeidet på 1970 og -80 tallet som ledd i denne prosessen, men på tross av fremskrittene fortsetter problemstillingen å øke i størrelse og kompleksitet (Bastian et al., 2010). For videre

utvikling på dette området kan det å forenkle prosessen med å utarbeide systematiske oversiktsartikler være et aktuelt tiltak. Det er likevel viktig at en forenkling av prosessen ikke går på bekostning av validiteten til studien, da dette er en av styrkene ved systematiske oversiktsartikler i dag (Schünemann & Moja, 2015).

3.1.1.1.. Hvorfor systematiske oversiktsartikler?

Dersom en ikke har kompetanse eller tid til å vurdere forskningen er det særlig viktig å bruke kilder med forhåndsvurdert innhold (Jamtvedt et al., 2003). Utarbeiding av systematiske oversiktsartikler stiller krav til systematiske metoder for å identifisere, velge ut, vurdere, sammenstille og gradere data fra empirisk forskning (Jamtvedt et al., 2003). Dette er, på tross av at det er tidkrevende, en effektiv forskningsmetode da det er mindre tidkrevende og mer kostnadseffektivt enn å starte nye enkeltstudier (Mulrow, 1994). Noen av fordelene ved å benytte oppsummert forskning som en kilde til informasjon er at det er effektivt, fører til mer presise resultater, minimerer risiko for systematiske skjevheter, hindrer muligheten for «plukking» av enkeltstudier, forkorter implementeringsprosessen og forbedrer reliabiliteten og nøyaktigheten av konklusjonene (Greenhalgh, 2019; Mulrow, 1994). Selv om enkeltstudier omhandler samme tema adresserer de ofte ulike definisjoner, populasjoner, intervensjoner og målemetoder. Systematiske oversiktsartikler av disse enkeltstudiene kan derfor bidra til å avgjøre om resultatene er generaliserbare, og eventuelle årsaker til heterogenitet kan identifiseres og føre til nye hypoteser om spesifikke subgrupper (Greenhalgh, 2019; Mulrow, 1994). Behovet for en grundig metode i utarbeidingen har resultert i utviklingen av en formell prosess for gjennomføringen av systematiske oversiktsartikler (Higgins & Green, 2008). En systematisk oversikt skal ha et klart formål, kriterier for inklusjon av artikler, en reproduserbar fremgangsmetode, en omfattende søkestrategi, kvalitetsvurdering av de inkluderte artiklene og en systematisk fremstilling av resultatene (Jamtvedt et al., 2003). På denne måten avdekkes det om det er konsensus rundt resultatene på tvers av de ulike enkeltstudiene (Mulrow, 1994). Transparens og reproduserbarhet sikres gjennom å dokumentere valgene som blir gjort underveis i prosessen (Higgins & Green, 2008).

3.1.1.2. Behov for hurtigere utarbeiding

Som beskrevet over er det mange fordeler og få bakdeler ved systematiske oversiktsartikler (Greenhalgh, 2019). Det er likevel noen forbehold som er viktig å være klar over. Kvaliteten på evidensen i en systematisk oversikt blir aldri bedre enn kvaliteten av enkeltstudiene den er basert på. Fullstendig identifikasjon av alle

relevante studier er viktig under utarbeiding av systematiske oversiktsartikler, men kan være utfordrende (Greenhalgh, 2019). Enkeltstudier som ville vært relevante for resultatene til den systematiske oversikten blir ofte ikke publisert, for eksempel på grunn av negative resultater, og selv studier som er publiser kan være vanskelige å finne (Dickersin et al., 1994). Det er derfor viktig med standardiserte og grundige metoder for alle stadiene ved utarbeidelsen av systematiske oversiktsartikler.

Innsamling, ekstraksjon og syntetisering av data på det medisinske fagfeltet er kjent for å være en tidkrevende prosess som ofte er utsatt for feil (Marshall, 2016). Det tar vanligvis forskere over et år å publisere en systematisk oversikt og 2,5 – 6,5 år før enkeltstudier implementeres i systematiske oversiktsartikler (Elliott et al., 2014; Jonnalagadda et al., 2015). I tillegg utdateres de systematiske oversiktene etter få år. Ved bruk av dagens metoder har Cochrane Collaboration ikke en gang klart å holde halvparten av sine systematiske oversiktsartikler oppdatert (Koch, 2006). En tidligere studie fant at mediantiden det tok før en systematisk oversiktsartikkel var utdatert i et forskningsfelt med relativt mange publikasjoner var 2.9 år, og noen artikler var utdaterte allerede før de var publisert (Shojania et al., 2007).

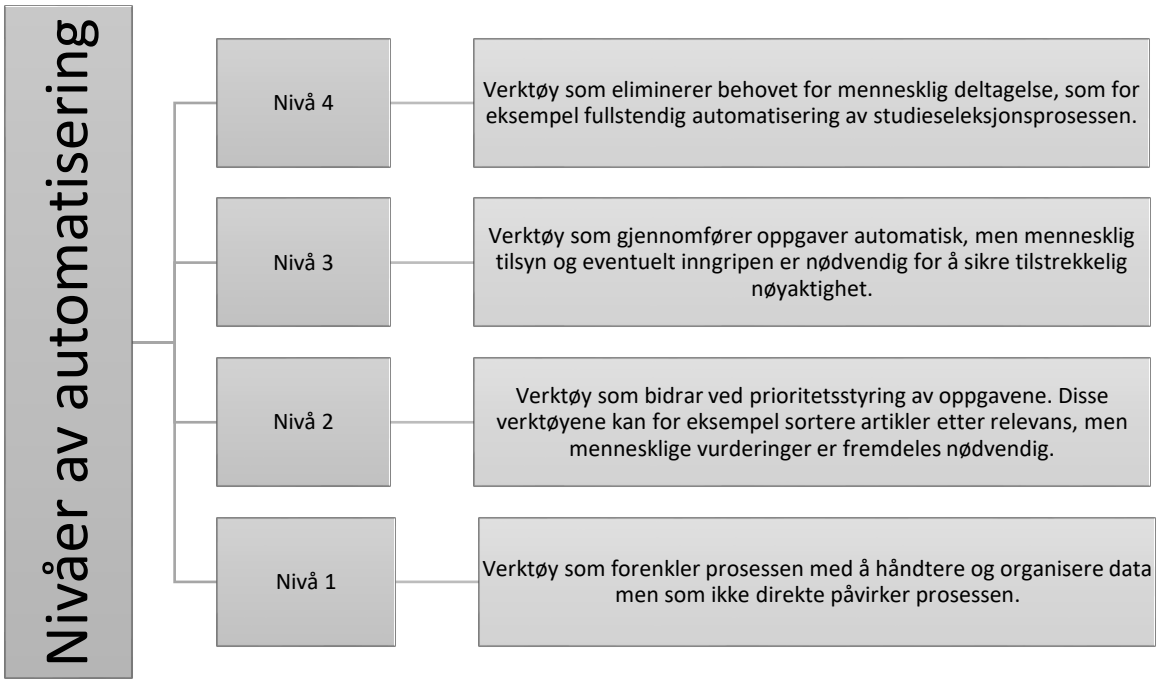
For å sikre at klinikere har tilgang til best mulig evidens er det viktig at forskning syntetiseres uten unødvendig forsinkelser. Vi må prioritere effektivt og redusere andelen unødvendig forskning fremover (Chalmers & Glasziou, 2009). Det har derfor blitt løftet frem et behov for hurtigere utvikling av systematiske oversiktsartikler (Schünemann & Moja, 2015). Det er likevel viktig at det utvikles adekvate metoder og prosedyrer for hurtigere utarbeiding av systematiske oversiktsartikler, da hurtighet ikke skal gå på bekostning av validiteten (Schünemann & Moja, 2015). Et alternativ for å oppnå dette er å utvikle gode metoder for kontinuerlig oppdatering av systematiske oversiktsartikler, som innebærer at artiklene oppdateres så fort ny forskning blir tilgjengelig (Elliott et al., 2017).

3.1.1.4. Automatisering av systematiske oversiktsartikler

Forventningene til bruk av KI innen det medisinske fagfeltet er skyhøye. KI kan potensielt assistere helsepersonell til å finne nyeste forskning, ta bedre kliniske beslutninger, og i noen tilfeller erstatte behovet for menneskelige vurderinger (Jiang et al., 2017). Disse egenskapene kan med fordel utnyttes under utarbeidelsen av systematiske oversiktsartikler, og en forskergruppe ved Bond University i Australia publiserte i 2020 en systematisk oversiktsartikkel som ved hjelp av ulike

programvarer ble utarbeidet på kun 2 uker (Clark et al., 2020). Det har tidligere blitt argumenterer for at menneskelige ressurser er mangelfulle og derfor kun bør benyttes i situasjoner hvor automatisering er umulig, upraktisk eller uønsket (Thomas et al., 2017). Hurtigere oppdatering av systematiske oversiktsartikler ved bruk av KI kan føre til personlig tilpasset behandling, bedre utfallsmål for pasientene og reduserte kostnader (Amann et al., 2022; Tsafnat et al., 2014).

Interessen for automatisering av utarbeidelsen av systematiske oversiktsartikler har blitt drevet av et behov for å fremskynde tilgjengeligheten av oppdatert forskningsbasert kunnskap (Ouzzani et al., 2016). KI, nærmere bestemt maskinlæring og naturlig språkbehandling, kan potensielt benyttes i flere stadier av utarbeidelsen av systematiske oversiktsartikler. For eksempel til litteratursøk, screening av sammendrag og ekstrahering av resultater (Cierco Jimenez et al., 2022; Marshall & Wallace, 2019; Tsafnat et al., 2014). Tidligere forskning på automatisering av systematiske oversiktsartikler angir at hovedformålet med automatiseringen er reduserte kostnader og hurtigere oppdatering ved at programvarer overtar oppgaver som i dag gjennomføres av mennesker (van Dinter et al., 2021). O´Connor et al (2019) presenterte i sin artikkel fire ulike nivåer av automatisering (Figur 2). Det er med dagens programvarer aktuelt med automatisering på nivå 1, 2 og potensielt 3. Graden av menneskelig inngripen som er nødvendig på nivå 3 avhenger av prestasjonen til maskinlæringsverktøyet. Forskere på feltet ser for seg at det på et tidspunkt potensielt kan utvikles et program på nivå 4 som automatisk henter ut relevante enkeltstudier, vurderer de, henter ut data, gjennomfører metaanalyser og produserer samt oppdaterer en systematisk oversiktsartikkel i sanntid (Tsafnat et al., 2014). Det finnes flere ulike typer KI, hvor noen er bedre egnet enn andre til å gjennomføre de ulike oppgavene.



Figur 2: Nivåer av automatisering, modifisert fra (O'Connor et al., 2019).

3.2. Kunstig intelligens

Som beskrevet i kapittel 3.1.1.4. kan flere av stadiene ved utarbeidelsen av systematiske oversiktsartikler automatiseres. Det er flere ulike former for KI som kan benyttes for å gjennomføre de ulike oppgavene. Og videre utvikling på dette område vil være helt essensielt for å nå målet om automatisering på nivå 3 og 4 (O'Connor et al., 2019). I dette kapittelet gis en kort innføring i KI og de undergruppene som er aktuelle å benytte ved studieseleksjon til systematiske oversiktsartikler.

3.3.1. Historisk bakgrunn, formål og definisjon

Begrepet KI har eksistert helt siden 1956 (Strümke, 2023). KI er ikke en spesifikk teknologi, men et samlebegrep som omfatter for eksempel maskinlæring, nevrale nettverk og naturlig språkbehandling (Davenport & Kalakota, 2019). Det finnes mange ulike definisjoner på KI, og disse endrer seg ofte og i takt med hva som er teknologisk mulig (Regjeringen, 2020). Det er vanskelig å definere intelligens, men EUs ekspertgruppe for KI definerer kunstig intelligens slik:

Kunstig intelligente systemer utfører handlinger, fysisk eller digitalt, basert på tolkning og behandling av strukturerte eller ustrukturerte data, i den hensikt å oppnå et gitt mål. Enkelte KI-systemer kan også tilpasse seg gjennom å analysere og ta hensyn til hvordan tidligere handlinger har påvirket omgivelsene (AI, 2019).

Utviklingen av KI har som mål å skape dataprogrammer som kan simulere menneskelig intelligens, og derfor gjennomføre oppgaver som hittil kun har vært mulig å gjennomføre for mennesker (Helsedirektoratet, 2022). En programvare basert på KI analyserer data, tar beslutninger og utfører handlinger basert på data fra for eksempel sensorer, kameraer, mikrofoner, trykkmålere eller data fra andre informasjonskilder (Regjeringen, 2020). Enkelte systemer har også en tilbakemeldingsmekanisme som gir algoritmene muligheten til å lære, enten fra egne erfaringer eller gjennom direkte tilbakemeldinger fra brukeren (Regjeringen, 2020). De siste årene har fagfeltet vært i stor utvikling, og programvarer basert på KI løser oppgaver stadig bedre (Strümke, 2023).

Flere typer KI kan benyttes til utarbeidelsen av systematiske oversiktsartikler. Avanserte metoder som dyp læring, tekstutvinning og nevrale nettverk kan potensielt utvikles og gi muligheter for full automatisering av systematiske oversiktsartikler i fremtiden (van Dinter et al., 2021). Det er likevel flere utfordringer som må adresseres før dette blir aktuelt, og med dagens teknologi er det mer aktuelt å semi-automatisere deler av prosessen med andre teknologier som maskinlæring og naturlig språkbehandling (Tsafnat et al., 2014; van Dinter et al., 2021).

3.3.2. Maskinlæring

I dag brukes begrepet KI hovedsakelig om maskinlæring, som går ut på at maskiner lærer å utføre oppgaver og løse problemer gjennom selvstendig prøving og feiling (Helsedirektoratet, 2022; Strümke, 2023). Dette er den mest lovende teknologien for videre utvikling i dag, og de fleste programvarer med KI baseres derfor på maskinlæring (Davenport & Kalakota, 2019; Regjeringen, 2020). Også i helsevesenet er maskinlæring en av de vanligste formene for KI, og kan ha stor nytteverdi både for klinikere og forskere (Davenport & Kalakota, 2019). Ledet av relevante kliniske spørsmål kan metoden for eksempel brukes til å forutsi hvilket behandlingsalternativ som vil være mest effektivt for den enkelte pasient ved å finne informasjon i massive mengder data (Jiang et al., 2017). Metoden kan også potensielt brukes til å avgjøre om en pasient vil pådra seg en bestemt sykdom, til bildediagnostikk og til forskning som for eksempel under utarbeidelsen av systematiske oversiktsartikler (Davenport & Kalakota, 2019; Tsafnat et al., 2014).

Maskiner har potensiale til å gjennomføre oppgaver i et hurtig tempo, men er fremdeles avhengig av menneskelig intelligens for å utforme oppgavene (Mueller & Massaron, 2021). Med utviklingen av maskinlæring, har det blitt mulig å lage programvarer som kan lære av data og eksempler fra den virkelige verden, uten at noen programmerer inn de spesifikke reglene for å løse et bestemt problem (Helsedirektoratet, 2022). Under utvikling av programvarer basert på maskinlæring, vil maskinlæringsalgoritmer bygge modeller basert på eksempeldata eller treningsdata som deretter brukes til å ta beslutninger (Regjeringen, 2020). Disse algoritmene er utviklet spesifikt for at maskinen selv skal lære seg å løse gitte problemer (Strümke, 2023). Maskinlæring kan også brukes til å analysere store datamengder og identifisere trender og mønstre som kan bidra i utvikling av medisiner og annen medisinsk forskning. Det er derfor svært aktuelt å benytte denne teknologien til systematiske oversiktsartikler (Davenport & Kalakota, 2019). Det finnes flere undergrupper av maskinlæring som kan være hensiktsmessige å benytte til ulike formål av utarbeidelsen, disse presenteres i de neste avsnittene.

3.3.2.1. Veiledet læring

Veiledet læring er den enkleste formen for maskinlæring og derfor den mest brukte, også når det kommer til utarbeidelsen av systematiske oversiktsartikler (Navarro et al., 2021; van Dinter et al., 2021). Utfordringen ved denne teknologien er at den krever data som inneholder et riktig svar for å trenes opp. Brukeren må da manuelt markere data for å gi tilbakemelding til programvaren på hva som er riktig svar – dette kalles annotering (Strümke, 2023). Eksempler på metoder for veiledet læring inkluderer tilfeldige beslutningsskoger, naïve bayes klassifiserere, gradientforsterkningsmaskiner, støttevektormaskiner og nevrale nettverk (Navarro et al., 2021). Veiledet læring kan være særlig aktuelt å benytte ved seleksjon av aktuelle studier til systematiske oversiktsartikler (Tsafnat et al., 2014). Flere av de aktuelle programvarene baserer seg på støttevektormaskiner, som er en maskinlæringsalgoritme som ved hjelp av annotering lærer å klassifisere data (Noble, 2006). Maskiner basert på slik veiledet trening har hovedsakelig to store begrensinger. Annotering av data er tidkrevende og kostbart, og maskinen får ved bruk av denne teknologien ikke anledning til å utforske verden fritt. Dette fører til at den ikke kan oppdage sammenhenger som vi ikke allerede er klar over. I enkelte tilfeller kan derfor andre former for maskinlæring være mer hensiktsmessig (Strümke, 2023).

3.3.2.2. Uveiledet læring og dyplæring

Dyplæring er en form for uveiledet læring som kan benyttes for å overkomme noen av utfordringene som følger med bruken av veiledet læring til utarbeidelsen av systematiske oversiktsartikler (Knafou et al., 2023). Dyplæringsalgoritmer kan trenes opp med ikke-veiledet læring, som er en form for automatisk dataanalyse hvor maskiner sorterer data for oss. Dataene inneholder, i motsetning til ved bruk av veiledet læring, ikke noe riktig svar og brukeren kan ikke bestemme hvordan maskinen skal sortere. Veiledet læring er derfor mest hensiktsmessig når man vet hva man ser etter, mens uveiledet læring kan benyttes for å finne nye mønstre og forklaringsmodeller (Strümke, 2023). K.R. Felizardo (2012; 2014; 2011) har publisert flere artikler som undersøker bruken av uveiledet læring til utarbeidelsen av systematiske oversiktsartikler istedenfor veiledet læring, som i større grad er utsatt for menneskelige feil grunnet repetitivt arbeid. En annen fordel med bruk av uveiledet læring for studieseleksjon til systematiske oversiktsartikler er at det ikke er behov for forhåndsvurdert treningsdata (Haddaway & Westgate, 2019). Bakdelen er at denne teknologien krever erfaring med maskinlæring og statistikk, og er derfor ikke et aktuelt verktøy for forskere med helsefaglig bakgrunn som ikke har erfaring med dette (van Dinter et al., 2021). En annen utfordring med denne teknologien er mangelen på transparens i enkelte løsninger basert på dyplæring. På områder der forklarbarhet er viktig, kan det derfor være bedre å velge en annen tilnærming enn dyplæring (Kommunal- og moderniseringsdepartementet, 2020).

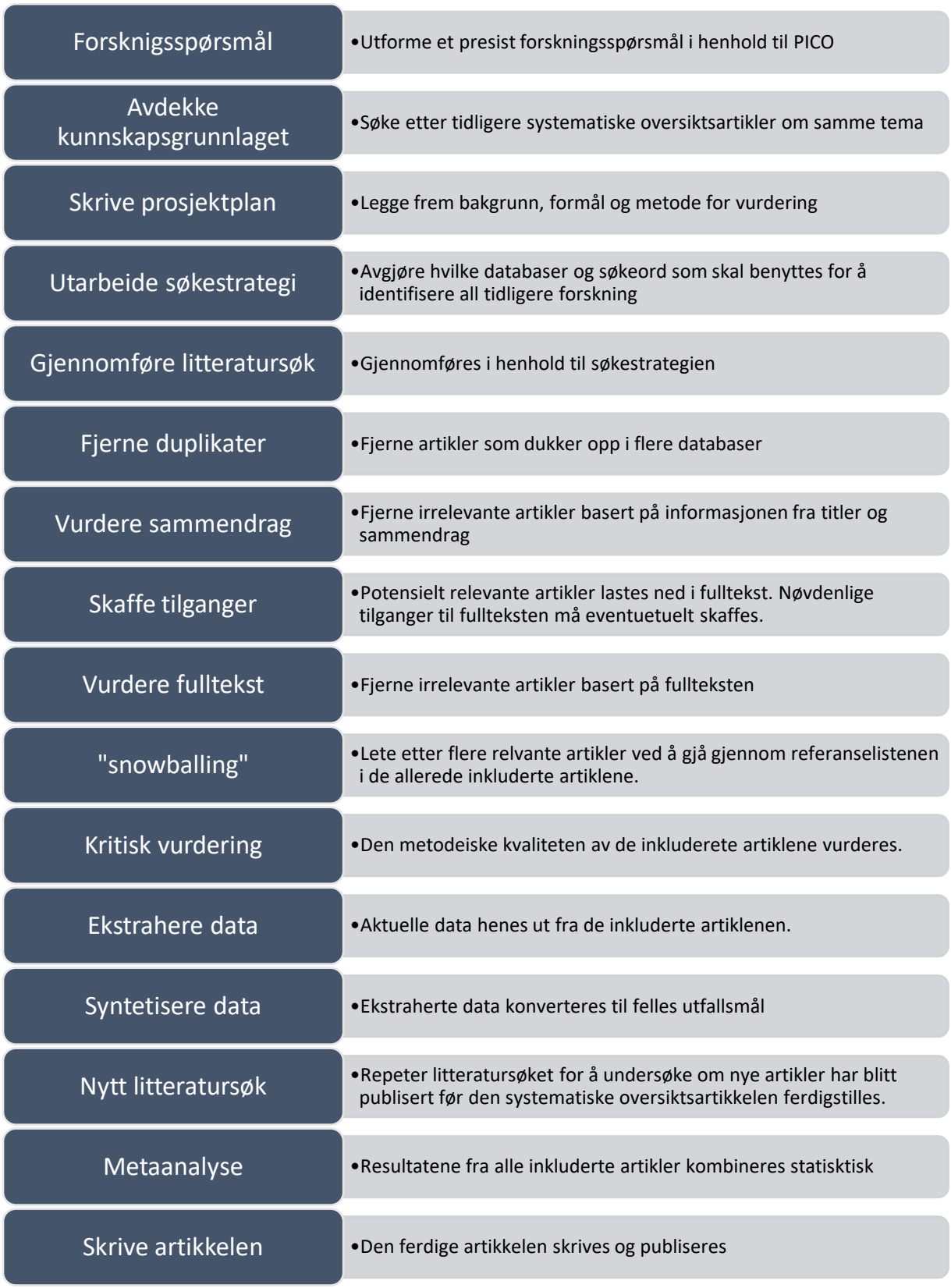
3.3.3. Naturlig språkbehandling (Natural language processing)

Denne formen for KI handler om å lære datamaskiner til å forstå og tolke menneskelig språk. Dette innebærer å analysere og tolke setninger, ordvalg, kontekst og meningsinnhold for å kunne gi relevante svar (Strümke, 2023). Naturlig språkbehandling kan brukes til alt fra å telle ord, generere setninger og til å klassifisere skriftlig data. Populære programvarer som ChatGPT er basert på denne formen for KI. Forskere har vært nysgjerrige på om programmet kan anvendes til utarbeidelsen av systematiske oversiktsartikler, men forsøk på dette har vist varierende resultater. En av grunnene til dette kan være at algoritmen i ChatGPT er designet for å gi svar basert på prediksjon og ikke ved å søke gjennom relevant litteratur (Qureshi et al., 2023). Selv om disse forsøkene viser et tydelig behov for videre utvikling brukes teknologien allerede i økende grad til å få økt innsikt i de enorme mengdene data på det medisinske fagfeltet, og til å utarbeide systematiske

oversiktsartikler (van Dinter et al., 2021). Tekstutvinning er en form for naturlig språkbehandling som har blitt spådd å kunne spille en stor rolle for utvikling av systematiske oversiktsartikler i tiden fremover (Lefebvre et al., 2013). Naturlig språkbehandling kan da benyttes til å hente ut relevant informasjon som videre kan bearbeides av maskinlæringsalgoritmer for å utforme systematiske oversiktsartikler (Blaizot et al., 2022).

3.4. Dagens metode for utarbeiding og potensialet for automatisering

Prosessen med å utarbeide en systematisk oversikt kan deles opp i fire til 16 steg, avhengig av hvordan man kategoriserer dette (Figur 3) (Higgins & Green, 2008; Tsafnat et al., 2014). Som beskrevet tidligere kan det være aktuelt å enten automatisere eller semi-automatisere flere av stegene ved hjelp av KI. Systematisk litteratursøk, studieseleksjon og ekstrahering av resultater er eksempler på steg som allerede med dagens teknologi er aktuelle å automatisere (Blaizot et al., 2022; Tsafnat et al., 2014). Andre stadier av utarbeidelsen, som det å undersøke behovet for en systematisk oversikt og selve skrivingen av rapporten, krever menneskelig kreativitet og innsikt og er derfor ikke aktuelle å automatisere med dagens teknologi (van Dinter et al., 2021). Dersom KI kan benyttes til å gjennomføre de tekniske stegene av utarbeidingen kan dette likevel være hensiktsmessig da det frigir mer tid til at forskerne kan gjennomføre de kreative oppgavene (Tsafnat et al., 2014). Det har blitt utviklet flere aktuelle verktøy og programvarer for utvikling av systematiske oversiktsartikler de siste årene, men ingen av de kan per i dag brukes til å gjennomføre alle stegene i prosessen. Det er derfor nødvendig å benytte ulike programvarer til ulike bruksområder (Ouzzani et al., 2016). I de neste avsnittene gis en kort beskrivelse av de ulike stadiene av utarbeidelsen sammen med mulighetene for automatisering, før det gås mer i dybden på steget for studieseleksjon og automatisering.



Figur 3: Oversikt over de ulike stadiene av utarbeidingen av en Systematisk oversiktsartikkel.

3.4.1. Formulere spørsmål og skrive prosjektplan

Formålet med en hver systematisk oversikt bør være klart formulert med et PICO-spørsmål som beskriver pasientgruppen, tiltaket, det eventuelle sammenligningsgrunnlaget og hvilke effekter eller utfallsmål man ønsker å sammenligne (Jamtvedt et al., 2003). Forskningsspørsmålet avgjør hvilket studiedesign som er mest hensiktsmessig for å besvare problemstillingen, da spørsmål om effekt, prognose, forekomst, årsak og erfaring har ulike foretrukne design (Helsebiblioteket, 2017). Forskningsspørsmålet legger også føringer for valg av søkeord og hvilke inklusjonskriterier som benyttes i utvelgelsen av studier, og det er derfor viktig med et presist spørsmål for å gjøre den videre prosessen enklere (Greenhalgh, 2019; Jamtvedt et al., 2003). Etter at formålet med studien er formulert i et forskningsspørsmål og eksklusjons- og inklusjonskriterier er utarbeider må det utarbeides en prosjektplan som beskriver hva den systematiske oversikten skal omhandle og hvordan den skal utarbeides (FHI, 2022).

Automatiske systemer kan bidra til å identifisere kunnskapshull, relevante pasientgrupper og problemstillinger, samt støtte forskerne i den kreative prosessen med å utforme relevante forskningsspørsmål. KI kan potensielt brukes til å forsikre at planen er konsekvent, objektiv og at metoden er hensiktsmessig for å svare på forskningsspørsmålet. Utforming av prosjektplanen er et av stegene som krever kreativitet og mye kunnskap om det aktuelle temaet. Formulering av spørsmål og utarbeiding av prosjektplanen er derfor blant de stegene hvor menneskelig kreativitet per i dag fremdeles er nødvendig (Tsafnat et al., 2014).

3.4.2. Systematisk litteratursøk

Det kan være vanskelig å finne alle aktuelle studier. Det er derfor viktig å utarbeide en god søkestrategi, og gjennomføre et systematisk litteratursøk for å sikre at alle potensielt aktuelle studier identifiseres. En god strategi sikrer også at identifiseringen av studier er så fri for systematiske skjevheter som mulig (Greenhalgh, 2019). Forfatterne av en systematisk oversikt bør søke i flere databaser, lete i referanselistene til hver enkelt av de inkluderte studiene og kontakte forfatterne av inkluderte artikler for å identifisere så mye relevant og oppdatert forskning som mulig (Jamtvedt et al., 2003). Det kan også være aktuelt å søke på publiserte prosjektplaner for å identifisere studier som ikke har blitt publisert. Et godt

gjennomført systematisk litteratursøk styrker tilliten vi kan ha til resultatene fra den systematiske oversikten (FHI, 2022).

Automatisering av litteratursøket kan bidra til å redusere tiden som brukes på å gjennomføre dette steget, føre til at flere relevante artikler identifiseres og øke presisjonen til søket (Tsafnat et al., 2014). Automatiske søk gjennomføres da regelmessig og forfatterne blir varslet når det blir publisert nye artikler som potensielt skal inkluderes i en systematisk oversiktsartikkel (Thomas et al., 2017). DistillerSR og RCT tagger er eksempler på programvarer som potensielt kan brukes for å gjennomføre systematiske litteratursøk (Cierco Jimenez et al., 2022). Det er likevel behov for videre forskning på naturlig språkbehandling for å utvikle algoritmer som kan forstå kliniske spørsmål og hente ut kontekst på en hensiktsmessig måte (Tsafnat et al., 2014).

3.4.3. Studieseleksjon

Etter gjennomføring av litteratursøket er neste steg i prosessen å selektere ut de artiklene som er relevante. Det vil da stort sett være et høyt antall artikler som ikke er aktuelle for inklusjon til den systematiske oversiktsartikkelen. Med dagens metoder må titler og sammendrag fra alle artiklene leses gjennom og vurderes av to uavhengige personer. Dette er svært tidkrevende og tidligere forskning på de eksisterende programvarene for automatisering av systematiske oversiktsartikler trekker frem at det med dagens teknologi er særlig relevant å automatisere studieseleksjonsprosessen. Derfor er det dette som utforskes nærmere i denne oppgaven. Studieseleksjonsprosessen, mulighetene for automatisering og aktuelle programvarer beskrives mer utdypende i kapittel 3.5.

3.4.4. Kritisk vurdering av inkluderte studier

Kvaliteten på evidensen i en systematisk oversikt blir som tidligere nevnt aldri bedre enn enkeltstudiene den er basert på. Det er derfor essensielt å kritisk vurdere kvaliteten av alle de inkluderte enkeltstudiene for å vurdere sannsynligheten for at metodiske feil har påvirket resultatene i den systematiske oversikten (Greenhalgh, 2019). Studiene vurderes både med tanke på intern og ekstern validitet. Det vurderes da både om forskningsspørsmålet er besvart på en måte som gjør at resultatene er mest mulig fri for systematiske skjevheter og i hvilken grad resultatene kan overføres til klinisk praksis (FHI, 2022). Standardiserte sjekklister og verktøy som GRADE (Grading of Recommendations Assessment, Development and Evaluation) kan

benyttes for å vurdere dette i systematiske oversiktsartikler om effekt av tiltak, diagnose og prognose (Goldet & Howick, 2013). Den metodiske kvaliteten av de inkluderte studiene vurderes også for å avgjøre hvor tungt resultatene skal vektes i den systematiske oversikten og metaanalysen. På denne måten kan vi sikre at en liten, men metodisk godt gjennomført, studie får den betydningen den fortjener sammenlignet med en større studie av lav metodisk kvalitet (Greenhalgh, 2019).

Tidligere studier har vist at ulike forfattere vurderer kvaliteten av enkeltstudiene ulikt (Lensen et al., 2014). Automatisering av dette steget kan potensielt bidra til at andelen menneskelige feil, samt den totale tidsbruken, reduseres. Marshall (2014) og Millard (2016) viser begge til lovende muligheter for kritisk vurdering med programvarer basert på maskinlæring. Det finnes imidlertid lite forskning på automatisering av dette steget enda, og mer forskning er dermed nødvendig før verktøyene kan tas i bruk (Jaspers et al., 2018).

3.4.5. Hente ut og sammenfatte data

Resultatene fra de inkluderte enkeltstudiene hentes ut og sammenfattes for hvert av de relevante utfallsmålene. Prosjektplanen spesifiserer hvilke data som er relevante å hente ut. Også denne prosessen gjennomføres av to personer for å kvalitetssikre prosessen. Sammenfatningen kan gjøres ved en deskriptiv syntese, eller så kan det i aktuelle tilfeller også gjennomføres en kvantitativ metaanalyse for å sammenfatte resultatene statistisk (FHI, 2022). Forfatterne må da avgjøre hvilke utfallsmål og måletidspunkt som er aktuelle å sammenfatte statistisk, før resultatene fremstilles (Greenhalgh, 2019). Det å hente ut og sammenfatte data er en av de mest tidkrevende stegene i prosessen, og derfor et av stegene med størst potensiale for effektivisering (Tsafnat et al., 2014)

Tidligere studier har undersøkt mulighetene for automatisering av dette steget, da selv delvis automatisering potensielt kan redusere arbeidsbelastningen ved utarbeiding av systematiske oversiktsartikler i betydelig grad (Tsafnat et al., 2014). Det er allerede standard prosedyre å benytte statistikkprogrammer til å gjennomføre sammenfatningen av data, men mye av jobben må fremdeles gjennomføres manuelt. Eksisterende programvarer som Colandr, RobotReviewer, DistillerSR og ExaCT kan brukes til å hente ut og sammenfatte data (Cierco Jimenez et al., 2022). Disse programvarene kan blant annet benytte naturlig språkbehandling for å markere relevant informasjon i teksten, eller maskinlæring for å assosiere ekstraherte data

med aktuelle utfallsmål (Hsu et al., 2012; Kiritchenko et al., 2010; Summerscales et al., 2009). Videre forskning på naturlig språkbehandling er nødvendig for å utvikle systemer som kan forstå tabeller, prosessere og sammenfatte all relevant informasjon (Tsafnat et al., 2014).

3.4.6. Presentere resultatene og skrive diskusjon

Etter at resultatene er hentet ut og sammenfattet må de beskrives og diskuteres i en rapport. Dette er en viktig del av å oppsummere forskning (FHI, 2022). Systematiske oversiktsartikler rapporters gjerne etter internasjonale standarder som PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) (Page et al., 2021). For å sikre at arbeidet og den endelige rapporten har god kvalitet skal den også fagfellevurderes og godkjennes før publisering (FHI, 2022).

Også til denne delen av utarbeidingen er det behov for videre forskning på naturlig språkbehandling for å forbedre programvarene og potensielt gjøre det mulig å automatiskere hele prosessen med å skrive en systematisk oversikt (Tsafnat et al., 2014). Dersom det utvikles programmer som kan skrive de systematiske oversiktene basert på prosjektplanene kan man potensielt spare flere år med arbeid (Tsafnat et al., 2014). RevMan-HAL er et program som kan fylle ut ferdige maler til Cochrane oversikter basert på metaanalyser. Forfatterne av den systematiske oversikten trenger da kun å redigere den ferdige teksten (Torres Torres & Adams, 2017).

3.5. Bruk av maskinlæring for studieseleksjon til systematiske oversiktsartikler

Hovedformålet med det systematiske litteratursøket er som nevnt i kapittel 3.4.2. å identifisere alle artikler som kan være relevante for inklusjon. Formålet med studieseleksjon, som er det påfølgende steget, er dermed å ekskludere artiklene som ikke er relevante (Tsafnat et al., 2014). Dette er ofte det mest kostbare og tidkrevende steget ved utarbeidelsen av en systematisk oversiktsartikkel, da det kan være mange falske positive treff etter et systematisk litteratursøk. For systematiske oversiktsartikler i det medisinske fagfeltet gjelder dette omtrent 97% av de identifiserte artiklene (Borah et al., 2017). For å sikre at alle relevante studier identifiseres og unngå seleksjonsskjevhet vurderes alle artikler fra det systematiske litteratursøket av to uavhengige personer, for om de er aktuelle for inklusjon til den systematiske oversikten (Ouzzani et al., 2016). Dette gjøres ved at begge personene først leser alle titler og sammendrag for å avgjøre hvilke artikler som skal leses i fulltekst før en endelig avgjørelse vedrørende inklusjon eller eksklusjon tas (FHI,

2022). Beslutningene vedrørende inklusjon eller eksklusjon av artiklene dokumenteres underveis for å sikre transparens, klarhet og sporbarhet av utvelgelsesprosessen. Dette styrker også kredibiliteten til den ferdige systematiske oversiktsartikkelen (Ouzzani et al., 2016). Dette er en tidkrevende prosess og det er blitt estimert at studieseleksjon basert på kun titler og sammendrag krever 10-20% av den totale tiden brukt på utarbeidelsen av en systematisk oversiktsartikkel (Haddaway & Westgate, 2019). Gjennomsnittlig brukes 332 timer på studieseleksjon til en enkelt systematisk oversiktsartikkel (Cohen et al., 2006). Tidsbruken kan derfor reduseres kraftig ved å øke presisjonen etter et systematisk litteratursøk (Cohen et al., 2006; Shemilt et al., 2016). Det er da viktig å benytte metoder som maksimerer sensitiviteten, slik at alle relevante studier identifiseres, samtidig som arbeidsbelastningen reduseres ved å øke spesifisiteten (Li et al., 2023).

Bruken av maskinlæring er derfor særlig relevant for å redusere tidsbruken og redusere kostnadene ved gjennomføring studieseleksjonsprosessen. Maskinlæringsverktøy kan benyttes for å semi-automatisere prosessen ved å identifisere relevante artikler til den systematiske oversiktsartikkelen basert på informasjonen fra titler og sammendrag (Cierco Jimenez et al., 2022; Marshall & Wallace, 2019). Maskinlæringsalgoritmer kan bruke tidligere avgjørelser til å predikere hvilke studier som er aktuelle for inklusjon (Gates et al., 2019). Særlig klassifiserere for randomiserte kontrollerte studier har vist lovende resultater, på grunn av tilgangen på store mengder treningsdata (Wallace et al., 2017). For å identifisere andre typer studier finnes det derimot ikke like gode klassifiserere på grunn av mangelen på treningsdata. Maskinlæringsverktøy kan likevel tas i bruk også når andre typer studier skal inkluderes, men verktøyene er da avhengig av å først lære av menneskelige beslutninger (Thomas et al., 2017). Maskinlæringsverktøyet kan basert på disse avgjørelsene brukes til å estimere sannsynligheten for om en artikkel bør inkluderes og rangere de gjenværende artiklene automatisk fra mest til minst relevant. Forskerne kan dermed potensielt identifisere studiene som skal inkluderes tidligere i prosessen (Marshall & Wallace, 2019). KI kan også potensielt benyttes til vurdering av artiklene i fulltekst, men mer forskning trengs for å vurdere reliabiliteten og mulighetene for integrering av dette (Tsafnat et al., 2014).

Semi-automatisering av studieseleksjonsprosessen gjennomføres i økende grad, og studier har vist at arbeidsbelastningen potensielt kan reduseres med 30-70% mot at 5% av de relevante studiene ikke identifiseres (O'Mara-Eves et al., 2015). Dersom sannsynligheten for å finne de relevante studiene på denne måten når et akseptabelt nivå kan prosessen også effektiviseres ved at kun en person vurderer artiklene ved hjelp av en programvare (FHI, 2022). Det er likevel fremdeles uklart hvordan og når forfatterne kan avgjøre at vurderingene gjort av programvarene er reliable nok til å benyttes (Gates et al., 2019). I en spørreundersøkelse gjennomført av Van Altena et al (2019) svarte 32% av respondentene at de bruker et automatisk verktøy ved utarbeiding av systematiske oversiktsartikler. Andre studier viser også at det er få forfattere av systematiske oversiktsartikler som benytter disse verktøyene, og forskning på de ulike programvarene er derfor nødvendig (Thomas, 2013).

3.5.1. Programvarer for studieseleksjon til systematiske oversiktsartikler

Det finnes allerede mange tilgjengelige programvarer for å effektivisere prosessen med vurdering og seleksjon av relevante studier (Tabell 1) (Cierco Jimenez et al., 2022; Harrison et al., 2020; Marshall & Wallace, 2019). Som vist i tabellen baserer disse programvarene seg på ulike former for KI, som naturlig språkbehandling, veiledet læring og dyplæring. Programvarer basert på naturlig språkbehandling kan automatisk fremheve setninger og ord, som forfatterne sannsynligvis kan benytte for å gjennomføre seleksjonen (Chung & Coiera, 2007; Thomas et al., 2011). Maskinlæring kan deretter estimere sannsynligheten for om en studie skal ekskluderes eller inkluderes, basert på de tidligere avgjørelsene til forfatteren ved bruk av veiledet læring (Cohen et al., 2012; Tsafnat et al., 2014; Wallace et al., 2010). Digitale verktøy som Covidence, Rayyan og EPPI Reviewer er ifølge FHI allerede relevante å bruke til denne prosessen under utarbeidelsen av systematiske oversiktsartikler (FHI, 2022). Eksempler på andre programvarer som trekkes frem i tidligere studier og som dermed kan være aktuelle er Abstrackr, RobotAnalyst og DistillerSR (Cierco Jimenez et al., 2022).

Tabell 1: Oversikt over tilgjengelige KI baserte programvarer for automatisering av studieseleksjonsprosessen.

Verktøy	*Teknologi*	*Tilgjengelighet*	*Tilgjengelig fra*
Abstrackr	SL	Gratis, nettbasert	http://abstrackr.cebm.brown.edu/account/login
*ASReview**	NLP, SL	Gratis, lastes ned	https://asreview.nl/
CADIMA	NLP	Gratis, nettbasert	https://www.cadima.info/
Cochrane RCT Classifier	SL	Tilgjengelig ved utarbeiding av systematiske oversiktsartikler for Cochrane	https://crsweb.cochrane.org/login.html
Colandr	NLP, SL	Gratis, nettbasert	https://www.colandrcommunity.com/
Covidence	NLP, SL	Mot betaling, nettbasert	https://www.covidence.org/
DistillerSR	NLP, SL	Mot betaling, nettbasert	https://www.distillersr.com/
DoCTER	NLP	Gratis, nettbasert	https://www.icf-docter.com/
Enago Read	NLP, SL	Gratis og mot betaling, lastes ned	https://www.read.enago.com/
EPPI-Reviewer	NLP	Mot betaling, lastes ned	https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=2914
IRIS.AI	NLP, SL	Gratis- og mot betaling, nettbasert	https://the.iris.ai/
PICO Portal	NLP	Mot betaling, nettbasert	https://picoportal.org/
RapidMiner	NLP, SL, USL	Mot betaling, lastes ned	http://rapid-i.com/
RAx – Enago read	NLP, SL	Mot betaling, nettbasert	https://www.read.enago.com/
Rayyan	NLP, SL	Gratis, nettbasert	https://www.rayyan.ai/
RevMan	NLP	Mot betaling, nettbasert	https://revman.cochrane.org/info
Research sceener	NLP, SL, DL	Gratis, nettbasert	https://researchscreener.com/
RobotAnalyst	SL	Ta kontakt med utvikler for tilgang, nettbasert	https://www.nactem.ac.uk/robotanalyst/
Screen4me	NLP	Tilgjengelig ved utarbeiding av systematiske oversiktsartikler for Cochrane	https://community.cochrane.org/sites/default/files/uploads/S4M_webinar_slides_Feb_2019.pdf
SR-Accelerator	NLP	Gratis, nettbasert	https://sr-accelerator.com/#/
SWIFT-Active screener	SL	Gratis, nettbasert	https://www.sciome.com/swift-activescreener/
SWIFT-Review	SL	Gratis, lastes ned	https://www.sciome.com/swift-review/
SyRF	NLP	Gratis, nettbasert	https://syrf.org.uk/
Sysrev	SL	Gratis, nettbasert	https://www.sysrev.com/

NLP = naturlig språkbehandling, SL = veiledet læring, DL = dyplæring, USL = uveiledet læring. Programvarer markert med en «*» krever kunnskap om programmering for å kunne benyttes.

3.5.2. Valg av programvare - Rayyan

Som nevnt finnes det flere aktuelle programvarer utviklet for å bidra ved studieseleksjon til systematiske oversiktsartikler. Til denne oppgaven ble maskinlæringsverktøyet Rayyan benyttet. Rayyan er utviklet for å semi-automatisere prosessen med å vurdere sammendrag for inklusjon eller eksklusjon, for å gjøre gjennomføringen mer effektiv. Under utvikling av programvaren ble det lagt fokus på brukervennlighet, verdien stjernerangeringen tilfører studieseleksjonsprosessen og nøyaktighet sammenlignet med manuelle metoder (Ouzzani et al., 2016). Rayyan ble valgt til denne oppgaven på bakgrunn av at det integrerte maskinlæringsverktøyet innehar hensiktsmessige funksjoner i forhold til oppgavene som skal utføres, er gratis, ikke krever kunnskap om dataprogrammering og at den er en av de oftest siterte programvarene (Cierco Jimenez et al., 2022). Særlig det at programvaren er gratis, nettbasert og ikke krever programmeringskunnskap var avgjørende for valget. Dette vil også kunne føre til at programmet ofte blir valgt av forskere innen helsefaglige områder.

Maskinlæringsverktøyet fungerer ved at Rayyan trekker ut enkeltord, par av ord (bigram) og tidligere beregnede MeSH-termer ved hjelp av naturlig språkbehandling, etter å ha fjernet stoppord, fra titler og sammendrag. Når brukere sorterer artikler i kategoriene ekskludert eller inkludert lærer støttevektormaskinen av egenskapene til disse artiklene, og bygger deretter en modell for å klassifisere de resterende. Artiklene gis da en stjernerangering fra 0.5 - 4.5 avhengig av hvor sannsynlig maskinlæringsverktøyet vurderer det er at studien skal inkluderes. Denne prosessen gjentas til det ikke er flere artikler å sortere eller til modellen ikke kan forbedres ytterligere (Ouzzani et al., 2016).

3.5.3. Tidligere forskning på Rayyan

Tidligere forskning på prestasjonen til det integrerte maskinlæringsverktøyet og funksjonaliteten til Rayyan har vist lovende resultater (Tabell 2). Flere forfattere av ulike systematiske oversiktsartikler har beskrevet programvaren som enkel å bruke, tidsbesparende og et nyttig verktøy for å bidra ved studieseleksjonen (Gaskins et al., 2021; Nascimento et al., 2021; Pinna et al., 2020; Rogers et al., 2020).

Ouzzani gjennomførte i 2016 en pilotstudie for å teste prestasjonen til støttevektormaskinen i Rayyan på 15 ulike systematiske oversiktsartikler. I denne studien ble 50% av studiene brukt til trening av programvaren og 50% til testing, før

resultatene ble sammenlignet med manuell sortering av artiklene. De trakk frem at programvaren er intuitiv og enkel å bruke. Brukere rapporterte i gjennomsnitt en tidsbesparelse på 40%. Forfatterne, som også har utviklet programvaren, konkluderte med at Rayyan er et nyttig verktøy som potensielt kan redusere arbeidsbelastningen ved utarbeiding av systematiske oversiktsartikler ved å gjøre prosessen med å inkludere og ekskludere studier mer tidseffektiv (Ouzzani et al., 2016).

Olofsson et al (2017) var de første uavhengige forskerne som evaluerte prestasjonen til maskinlæringsverktøyet i Rayyan ved studieseleksjon til seks ulike systematiske oversiktsartikler. Ved å sortere studiene etter stjernerangeringene kalkulert av maskinlæringsverktøyet i Rayyan fant de ut at medianverdien for sensitivitet var 60% etter at 25% av artiklene var manuelt vurdert, 95% etter at halvparten av artiklene var manuelt vurdert og 98% etter manuell vurdering av 75%. Det var stor spredning i verdiene etter vurdering av 25%, hvor sensitiviteten varierte fra 21% til 88% avhengig av hvilken systematisk oversiktsartikkel maskinlæringsverktøyet ble testet på. Ved nærmere retrospektiv undersøkelse oppdaget de imidlertid at studiene med lav risiko for feil ble identifisert tidlig i prosessen. De konkluderte derfor med at prestasjonen til maskinlæringsverktøyet Rayyan imøtekom forventningene og effektivt kan benyttes for å assistere forfatterne ved gjennomføring av studieseleksjonsprosessen (Olofsson et al., 2017).

Valizadeh (2022) utforsket effektiviteten av Rayyan for screening av sammendrag av studier til tre systematiske litteraturstudier om diagnostisk nøyaktighet. Resultatene indikerte at et Rayyan er et effektivt og nøyaktig verktøy for screening av sammendrag, og at verktøyet potensielt kan redusere tidsbruken og arbeidsbelastningen for forskere som utarbeider systematiske oversiktsartikler. Ved en grenseverdi på <2.5 stjerner for eksklusjon oppnådde maskinlæringsverktøyet en gjennomsnittlig sensitivitet på 97-99% og spesifisitet på 19-58%. Rayyan hadde dermed relativt god evne til å ekskludere irrelevante artikler, men presterte ikke like godt når det kom til å identifisere relevante artikler. De påpekte at Rayyan dermed har begrensninger, og at det derfor likevel er nødvendig med menneskelig vurdering av en andel av sammendragene (Valizadeh et al., 2022).

Den tidligere nevnte studien utført av Li et al (2023) undersøkte hvordan tekstutvinning påvirket sensitiviteten og spesifisiteten ved studieseleksjon til en systematisk oversikt, blant annet ved bruk av Rayyan. De vurderte studier med 2.5

stjerner eller mer som aktuelle for inklusjon, da denne grenseverdien av Valizadeh ble vist å maksimere sensitiviteten. Resultatene viste gode sensitivitetsverdier, men lav spesifisitet. Dersom studieseleksjonen ble gjennomført automatisk, uten menneskelig vurdering av stjernerangeringene, var sensitiviteten 100% og spesifisiteten 26.7%. Dette vil si at alle relevante studier ble identifisert, men at mange irrelevante studier ble vurdert som aktuelle for inklusjon. De argumenterte derfor for at arbeidsbelastningen potensielt kan reduseres ved bruk av Rayyan, men menneskelig assistanse er nødvendig for å øke spesifisiteten (Li et al., 2023).

Felles for disse studiene er at alle trakk frem behovet for ytterligere forskning på programvaren. Blant annet for å avgjøre betydningen av de studiene som potensielt mistes ved å benytte Rayyan for å vurdere sammendrag og for å vurdere brukertilfredshet, brukervennlighet, potensiell reduksjon i arbeidsbelastning, samt risikoer og fordeler assosiert med denne metoden for studieseleksjon (Chai et al., 2021; Tsou et al., 2020; Valizadeh et al., 2022). Det er også behov for å undersøke effekten av verktøyet ved seleksjon av studier med ulike studiedesign, ulik presisjon på søkestrategien, testing etter at en lavere prosentandel av studiene er menneskelig vurdert og om det kan settes en grenseverdi for automatisk ekskludering av artikler (Olofsson et al., 2017; Valizadeh et al., 2022). Li (2023) presiserte at resultatene fra deres studie kan ha blitt påvirket av høy punktprevalens og at mer presise estimater av sensitivitet og spesifisitet kunne blitt oppnådd med et større datasett.

Tabell 2: Tidligere forskning på prestasjonen til maskinlæringsverktøyet og andre funksjoner i Rayyan.

Referanse	Formål	Metode	Tema	PR	Resultat
Prestasjonen til maskinlæringsverktøyet					
(Ouzzani et al., 2016)	Betatesting av prestasjonen til programvaren på 15 ulike systematiske oversiktsartikler, inkludert en brukerundersøkelse.	50% av artiklene ble brukt til trening og 50% til testing.	15 ulike medikamenter.	0.5-21.7%	49% reduksjon i tidsbruk ved en sensitivitet på 95%,
(Olofsson et al., 2017)	Undersøke prestasjonen til Rayyan vedrørende studieseleksjon til 6 systematiske oversiktsartikler, inkludert en brukerundersøkelse.	Antallet relevante studier identifisert av Rayyan ble vurdert etter at 25%, 50% og 75% av artiklene var sortert.	Foster-diagnostikk. Artrose. Hjerte- og karsykdom. Angst. Sårstell. Traume mot perineum.	1%-18%.	SEN på 21-88% etter at 25% var vurdert. SEN på 86-99% etter at 50% var vurdert. SEN på 89-100% etter at 75% var vurdert.
(Valizadeh et al., 2022)	Å evaluere prestasjonen til det integrerte maskinlæringsverktøyet i Rayyan.	Søkeresultatet til tre ulike systematiske oversiktsartikler ble manuelt vurdert i fire bolker. For hver bolk ble Rayyan brukt til å vurdere de resterende artiklene. Prestasjonen til Rayyan ble vurdert ved å kalkulere sensitiviteten, spesifisiteten, NPV, PPV og F1-skår ved grenseverdier på ≤2.5 og <2.5 stjerner for eksklusjon.	Diagnostisk nøyaktighet ved bruk av maskinlærings-algoritmer til tolkning av cerebrale MR bilder og EEG undersøkelse.	12-80%	Gjennomsnittsresultat ved grenseverdi <2.5 for eksklusjon: SEN = 97-99% SPE = 19-58% PPV = 22-31% NPV = 98-100%

					Gjennomsnitts-resultat ved grenseverdi ≤2.5 for eksklusjon: SEN = 1-30% SPE = 100% PPV = 86-97% NPV = 56-88%
(Scherhag & Burgard, 2023)	En retrospektiv analyse av potensiell reduksjon i arbeidsbelastning og ulike stoppkriterier ved bruk av Rayyan og ASReview.	Sammenlignet resultatene ved 7 ulike stoppkriterier. 50 /100 påfølgende artikler vurdert som irrelevante Stoppe etter vurdering av 25%, 50% og 75% av artiklene Stoppe på en grenseverdi <2.5 stjerner Stoppe når det er estimert at 95% av relevante artikler er identifisert.	Medikamenter.	0.5%-27%	Rayyan reduserte arbeidsbelastningen med 20-33%. ASReview presterte bedre med opptil 54%. Vurdering av artikler i 25% bolker var den mest reliable fremgangsmetoden.
(Li et al., 2023)	Å evaluere påvirkningen tekstutvinning har på sensitiviteten og spesifisiteten ved studieseleksjon til systematiske oversiktsartikler.	Prestasjonen ble vurdert for vurdering av 500 artikler til en systematisk oversikt. 200 artikler ble brukt til trening av maskinlærings-algoritmen, hvorav 15 var relevante. Studier med 2.5 stjerner eller mer ble vurdert som aktuelle for inklusjon	Komplikasjoner ved graviditet	10%	Gjennomsnittlige sensitivitets- og spesifisitetsverdier: SEN: 97% SPE: 95.3%

Brukervennlighet eller andre funksjoner					
(Cleo et al., 2019)	Undersøke brukervennligheten og akseptabiliteten til Covidence, SRAHelper for EndNote, Rayyan og RobotAnalyst.	Tre forfattere av systematiske oversiktsartikler testet programvarene og svarte på en brukerundersøkelse.	Ikke aktuelt.	Ikke aktuelt.	SRA-Helper for EndNote fikk 28/30 poeng, Covidene fikk 23/30, Rayyan fikk 25/30 og RobotAnalyst 22/30.
(Harrison et al., 2020)	Identifisere, beskrive og evaluere brukervennligheten av tilgjengelige programvarer som assisterer seleksjonsprosessen ved utarbeiding av systematiske oversiktsartikler.	Studien hadde fire stadier. En søkestrategi ble utvikler for å identifisere programvarene, kriterier for inklusjon ble utformet og en funksjonsanalyse og brukerundersøkelse ble gjennomført.	Ikke nevnt.	Ikke aktuelt.	15 programvarer møtte inklusjonskriteriene. Rayyan fikk den høyeste poengsummen etter funksjonsanalysen, men var tredje beste alternativ i følge brukerundersøkelsen (slått av Covidence og Abstrackr) .
(Yu et al., 2022)	Undersøke prestasjonen og brukervennligheten til både nett- og mobilversjonen av Rayyan.	Data ble samlet inn gjennom spørreskjemaer, data fra Rayyan og «System Usability Scale»	Ikke nevnt	Ikke nevnt	Ingen forskjell i prestasjonen til nett- og mobilversjonen.

3.5.4. Tidligere forskning på prestasjonen til andre programvarer

Forskning på prestasjonen til andre maskinlæringsverktøy for studieseleksjon basert på titler og sammendrag til systematiske oversiktsartikler har også vist lovende resultater (Chai et al., 2021; Tsou et al., 2020). Gates et al gjennomførte i 2019 en studie hvor de undersøkte fordelene og risikoene ved å benytte maskinlæringsverktøyene Abstrackr, DistillerSR og RobotAnalyst for å semi-automatisere studieseleksjonen. Maskinlæringsalgoritmene ble trent opp ved hjelp av manuell vurdering av 200 artikler. Resultatene viste at alle tre verktøyene kan bidra til å redusere arbeidsbelastningen med 85-99%. Andelen relevante studier som ble mistet ved helautomatisering av oppgaven var 5% ved bruk av Abstrackr, 97% ved bruk av DistillerSR og 70% ved bruk av RobotAnalyst. Ved semi-automatisering var resultatene 1%, 2% og 2%, men da med en betydelig mindre reduksjon i arbeidsbelastningen på 35-49%. Forfatterne konkluderte med at Abstrackr med fordel kan benyttes, men at automatisering medfører en risiko for at ikke alle relevante studier identifiseres. Videre presiserte de at videre forskning er nødvendig for å fastslå hvordan maskinlæring best kan benyttes for å redusere arbeidsbelastningen og identifisere hvilke arbeidsoppgaver som er best egnet for automatisering (Gates et al., 2019).

I artikkelen fra Tsou (2020) ble de to maskinlæringsverktøyene Abstrackr og EPPI-Reviewer, som begge kan benyttes for vurdering av titler og sammendrag, beskrevet som nyttige verktøy for å redusere tidsbruken og arbeidsbelastningen forbundet med studieseleksjonsprosessen. De undersøkte hvor godt maskinlæringsverktøyene presterte ved studieseleksjon til ni ulike systematiske oversiktsartikler. EPPI-Reviewer viste bedre resultater enn Abstrackr både på nøyaktighet og effektivitet, selv om resultatene i stor grad varierte avhengig av til hvilken systematisk oversikt programvarene ble benyttet. Medianverdien for andelen artikler som måtte manuelt sorteres før programvarene oppnådde sensitivitet på 100% var 95.6% ved bruk av Abstrackr og 91.3% ved bruk av EPPI-Reviewer. Ved seleksjon av studier til tre store systematiske oversiktsartikler oppnådde Abstrackr en potensiell reduksjon i arbeidsbelastningen på 4-49%, mens EPPI-Reviewer potensielt reduserte den med 9-60%. Også ved bruk til mindre systematiske oversiktsartikler viste EPPI-Reviewer mest lovende prestasjon, men forskjellen var stort sett ikke statistisk signifikant og begge verktøyene kan potensielt benyttes (Tsou et al., 2020).

Harrison et al identifiserte i sin studie fra 2020 15 ulike maskinlæringsverktøy som kan benyttes ved studieseleksjon til systematiske oversiktsartikler. De undersøkte hvilke funksjoner de ulike programvarene har, og gjennomførte en brukerundersøkelse for å kartlegge brukeropplevelsen ved de seks verktøyene med høyest skår på funksjonsanalysen. Av disse trekkes Covidence, Rayyan og Abstrackr frem som gode alternativer, mens Colandr, DRAGON og EPPI-Reviewer konsekvent presterte dårligere enn de tre andre alternativer. Covidence fikk den høyeste poengsummen, men krever betaling for å kunne benyttes. Rayyan ble trukket frem som et brukervennlig gratisalternativ, og alle respondentene ville benyttet denne programvaren igjen (Harrison et al., 2020).

En annen programvare som benytter mer avansert teknologi og som har vist lovende prestasjon ved vurdering av sammendrag til systematiske oversiktsartikler er Research Screener. Resultatene fra en artikkel publisert i 2021 (Chai et al.) viste at denne programvaren redusert arbeidsbelastningen med mellom 60 og 96% ved bruk til ni ulike systematiske oversiktsartikler og to sonderende oversikter. Programvaren oppnådde sensitivitet på 100% etter vurdering av gjennomsnittlig 10.6% av artiklene til de systematiske oversiktsartiklene. Medianverdien var 6%. De konkluderte på bakgrunn av dette med at forskere ved å menneskelig vurdere 50% av artiklene med høy sikkerhet kan anta at 100% av alle relevante artikler er identifisert, noe som drastisk kan redusere tidsbruken og dermed kostnadene. (Chai et al., 2021)

Li et al sammenlignet i 2023 fem ulike metoder for studieseleksjon, hvor studiene ble vurdert av en til to forfattere med og uten bruk av maskinlæringsverktøy. De sammenlignet også reduksjonen i tidsbruk, samt sensitiviteten og spesifisiteten til maskinlæringsverktøyene Rayyan, Abstrackr og SWIFT-Review. De fant at bruk av et maskinlæringsverktøy resulterte i redusert tidsbruk og forbedret spesifisitet sammenlignet med konvensjonelle metoder, uten at dette gikk på bekostning av sensitiviteten. Ved automatisk eksklusjon av artiklene maskinlæringsverktøyene vurderte som irrelevante, oppnådde både Rayyan, Abstrackr og SWIFT-review en gjennomsnittlig sensitivitet på 100%. Mens spesifisiteten var gjennomsnittlig 26.7% for Rayyan, 82.2% for Abstrackr og 86.2% for SWIFT-Review. (Li et al., 2023).

4. Metode

4.1. Studiedesign

Til denne oppgaven ble det gjennomført en metodestudie for å undersøke prestasjonen til det implementerte maskinlæringsverktøyet i Rayyan for semi-automatisering av studieseleksjon til en systematisk oversiktsartikkel på prognostiske modeller for utfall etter degenerativ ryggkirurgi. For å undersøke om maskinlæringsverktøyet kan levere nøyaktige resultater og dermed brukes for å optimalisere og effektivisere utarbeidelsen av en systematisk oversiktsartikkel ble evnen til å identifisere artikler som er relevante for inklusjon sammenlignet med menneskelige vurderinger som referansestandard.

4.1. Systematisk oversikt

Den systematiske oversikten utføres av en forskningsgruppe ved senter for intelligent muskel-skjeletthelse ved OsloMet (Center for Intelligent Musculoskeletal health, 2023). I et av de pågående prosjektene, AID-spine, utarbeides det en stor systematiske oversiktsartikkel som omhandler prognostiske modeller for utfall etter degenerativ ryggkirurgi. Oversiktsartikkelen er fremdeles under utarbeiding, men litteratursøk og studieseleksjon er ferdigstilt. For nærmere beskrivelse av metode er protokollen publisert i det internasjonale registeret for systematiske oversiktsartikler (PROSPERO) og tilgjengelig online (PROSPERO 2022 CRD42022370499). Artikkelen skal ifølge protokollen ta for seg følgende tre forskningsspørsmål:

- Hva er evidensgrunnlaget rundt tilgjengelige prognostiske modeller for å forutsi vellykkede/ dårlige helseutfall hos personer med korsryggsmerter som mottar kirurgisk behandling?
- Hvordan er kvaliteten på tilgjengelige prognostiske modeller, og i hvilken grad forklarer de vellykkede / dårlige helseutfall hos personer med korsryggsmerter som får kirurgisk behandling?
- I hvilken grad er de identifiserte prognostiske modellene eksternt validert?

Det primære utfallsmålet i den systematiske oversikten er vellykkede eller mislykkede utfallsmål i etterkant av ryggkirurgi, sett i betraktning av smerteintensitet eller grad av funksjonsnedsettelse. Studier som undersøkte andre relevante parametere som dødelighet, reoperasjoner og lengde på sykehusopphold ble også inkludert.

4.1.2. Søkestrategi til den systematiske oversikten

Det ble gjennomført søk i databasene EMBASE, MEDLINE og CINAHL, hvor det ble brukt en kombinasjon av søkeord relatert til korsryggsmerter, prognostiske modeller og ryggkirurgi. Det ble ikke lagt noen restriksjoner for språk eller publiseringstidspunkt, og både prospektive kohortstudier, randomiserte kontrollerte studier og registerbaserte studier var aktuelle for inklusjon. Fullstendig søkestrategi er tilgjengelig i protokollen (PROSPERO 2022 CRD42022370499).

4.1.3. Inklusjon- og eksklusjonskriterier til den systematiske oversikten

Ved inklusjon av studier til denne oppgaven fikk jeg tilgang til og tok utgangspunkt i de menneskelige vurderingene som ble gjort ved inklusjon og eksklusjon til den systematiske oversikten. Samme studier har derfor blitt inkludert og ekskludert. Følgende studier ble vurdert som relevante for inklusjon:

- Studier som rekrutterte voksne pasienter av begge kjønn med korsryggsmerter, med eller uten utstrålende smerter, som gjennomgikk kirurgisk behandling.
- Studier som utviklet eller validerte prognostiske modeller for vellykkede eller mislykkede resultater etter kirurgisk behandling.
- Alle studiedesign med pasientoppfølging ble inkludert. Dette inkluderer for eksempel prospektive kohortstudier, randomiserte kontrollerte studier og registerbaserte studier.
- Studier ble inkludert uavhengig av språk og publikasjonsdato.

Følgende studier ble ekskludert:

- Studier som undersøkte smerter relatert til alvorlig spinal patologi

4.1.4. Manuell studieseleksjon

Studieseleksjonen ble gjennomført av tre forfattere med erfaring fra utarbeiding av flere tidligere systematiske oversiktsartikler. Alle titler og sammendrag til artiklene fra det systematiske litteratursøket ble vurdert av minst to forfattere, som gjennomførte vurderingen uavhengig av hverandre. Det var totalt 7994 antall treff etter litteratursøket. Totalt ble 127 artikler vurdert som aktuelle for inklusjon basert på tittel og sammendrag, og dermed vurdert i fulltekst. Eventuelle konflikter i studieseleksjon ble løst ved konsensus mellom de tre forfatterne. Til slutt ble referanselister sjekket for ytterligere relevante artikler, men ingen artikler ble inkludert herfra.

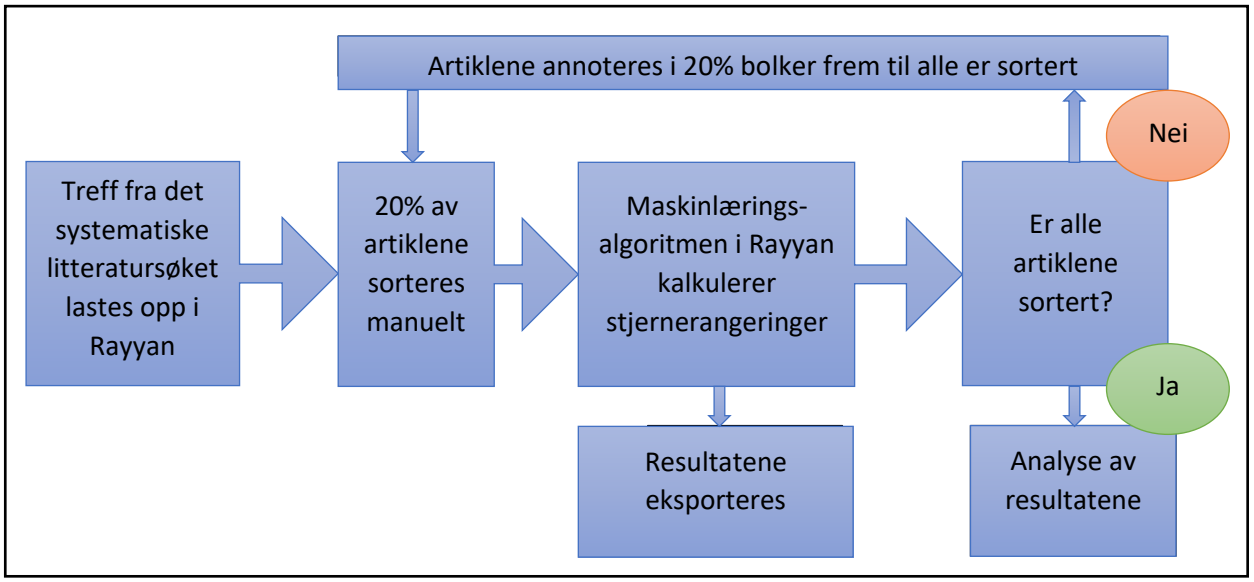
4.2. Prosedyre for datainnsamling og analyse

Maskinlæringsverktøyets evne til å identifisere artikler som er relevante for inklusjon til en systematisk oversikt ble undersøkt. Dette ble gjennomført ved at artiklene ble manuelt sortert til kategoriene inkludert eller ekskludert i 20% nivåer (fra 20% til 80%). Sorteringen ble gjennomført basert på de menneskelige vurderingene som ble gjort ved den manuelle studieseleksjonen til den systematiske oversiktsartikkelen. Ved hvert nivå ble det integrerte maskinlæringsverktøyet brukt til å rangere de resterende artiklene etter relevans. Stjernerangeringene som ble generert av Rayyan for de resterende artiklene er intervensjonen som ble vurdert, mens avgjørelsene som ble gjort av forskerne som vurderte treffene manuelt er sammenligningsgrunnlaget.

4.2.1. Fremgangsmåte:

Følgende fremgangsmåte ble benyttet for å generere stjernerangeringene som danner grunnlaget for evalueringen av prestasjonen til maskinlæringsverktøyet. Fremgangsmåten er også illustrert i figur 4.

1. Alle treff fra det systematiske søket ble importert til Rayyan.
2. Artiklene ble sortert alfabetisk etter forfatter.
3. De første 20% av artiklene ble gjennomgått og manuelt sortert til kategoriene «inkludert» eller «ekskludert».
4. Algoritmen i Rayyan vurderte relevansen til de resterende artiklene og gav de en stjernerangering fra 0.5 (minst relevant) til 4.5 stjerner (mest relevant).
5. Artiklene ble deretter sortert etter stjernerangering og hver enkelt ble gitt en markering tilsvarende stjernerangeringen.
6. Stjernerangeringene ble deretter eksportert fra programvaren som en CSV fil.
7. Etter at de første resultatene var hentet ut ble artiklene igjen sortert alfabetisk etter forfatter.
8. Steg 3-7 ble repetert for hvert 20% nivå. Det vil si at nye stjernerangeringer ble hentet ut etter at 1599, 3198, 4796 og 6395 studier var manuelt sortert.
9. CSV filene for hvert 20% nivå ble deretter importert til et statistisk analyseprogram for videre bearbeiding og analyser.



Figur 4: Illustrasjon av fremgangsmåten som ble benyttet for å generere stjernerangeringene i Rayyan.

4.2.2. Utfallsmål og analyser av resultatene

To ulike grenseverdier for eksklusjon/inklusjon av artiklene ble benyttet: < 2.5 stjerner for eksklusjon (artikler med en rangering på 2.5, 3.5 og 4.5 ble vurdert som relevante for inklusjon) og ≤ 2.5 stjerner for eksklusjon (artikler med en rangering på 3.5 og 4.5 ble vurdert som relevante for inklusjon). Disse to grenseverdiene ble valgt på bakgrunn av resultatene fra tidligere studier som indikerte at dette vil gi de mest balanserte resultatene for både sensitivitets- og spesifisitetsverdier (Valizadeh et al., 2022). I tillegg kan det tenkes at artiklene som fikk 2.5 stjerner var mest utfordrende å vurdere både for algoritmen og for forfatterne av den systematiske oversiktsartikkelen. Det er derfor interessant å undersøke prestasjonen til maskinlæringsverktøyet ved begge disse grenseverdiene for eksklusjon. Krysstabeller ble utarbeidet for hvert 20% nivå i seleksjonsprosessen for å sammenligne frekvensfordelingen mellom inkluderte og ekskluderte artikler basert på avgjørelsene fra maskinlæringsverktøyet sammenlignet med menneskelige vurderinger. Statistikkprogrammet SPSS versjon 27 ble benyttet for gjennomføring av alle analyser. Krysstabellene ble benyttet til å lage forvirringsmatriser som igjen ble benyttet for å kalkulere sensitivitet, spesifisitet, positiv prediktiv verdi (PPV) og negativ prediktiv verdi (NPV) for hvert nivå av studieseleksjonsprosessen (Tabell 3).

Tabell 3: Eksempel på forvirringsmatrise med formler for utregning.

		Manuell vurdering		
		Inkludert	Ekskludert	
Rayyan	≥ 2.5 stjerner (inkludert)	tp	fp	PPV = tp/(tp+fp)
	< 2.5 stjerner (ekskludert)	fn	tn	NPV = tn/(tn+fn)
		SEN = tp/(tp+fn)	SPE = tn/(tn+fp)	PR = (tp+fn)/(tp+fn+fp+tn)

Forkortelser: PPV = positiv prediktiv verdi, NPV = negativ prediktiv verdi, SEN = sensitivitet, SPE = spesifisitet, PR = punktprevalens, True positive (tp) = Inkludert av Rayyan og ved menneskelige vurderinger, False positive (fp) = Inkludert av Rayyan, men ekskludert ved menneskelige vurderinger, True negative (tn) = Ekskludert av Rayyan og ved menneskelige vurderinger, False negative (fn) = Ekskludert av Rayyan, men inkludert ved menneskelige vurderinger.

Sensitiviteten er andelen artikler som ble vurdert til å være relevante for inklusjon av Rayyan blant de artiklene som faktisk ble inkludert ved menneskelige vurderinger. Sensitiviteten regnes ut ved å dele antallet artikler som ble korrekt identifisert som relevante av Rayyan på det totale antallet relevante artikler Spesifisiteten er andelen artikler som ble vurdert av Rayyan til å ikke være aktuell for inklusjon blant alle artiklene som ble ekskludert. Denne ble regnet ut ved å dele antallet artikler som ble korrekt identifisert som ikke relevante av Rayyan på det totale antallet ikke relevante artikler. PPV er sannsynligheten for at når artikkelen vurderes som aktuell for inklusjon av Rayyan, så skal den faktisk inkluderes. PPV regnes ut ved at alle artiklene som ble korrekt identifisert som relevant av Rayyan deles på det totale antallet artikler som ble vurdert som relevante for inklusjon. NPV er sannsynligheten for at når artikkelen vurderes som ikke aktuell for inklusjon av Rayyan, så skal den faktisk ekskluderes. NPV regnes ut ved å dele antallet artikler som ble korrekt identifisert som ikke relevante på det totalet antallet artikler Rayyan vurderte som ikke relevante for inklusjon. Fordi PPV og NPV er avhengig av «prevalensen» av relevante artikler, som igjen avhenger av sensitiviteten til søkestrategien, ble også punktprevalensen kalkulert for hvert nivå. (Tabell 3). Til slutt ble relevante data brukt til å utarbeide grafer for å bedre illustrere resultatet.

4.3. Etiske vurderinger

I all forskning er det viktig å underordne seg etiske prinsipper og juridiske retningslinjer (Johannessen et al., 2010). Etikk er også et viktig tema under utvikling og bruk av programvarer basert på KI (Strümke, 2023). Jeg måtte derfor ta både EUs etiske retningslinjer for KI og vår egen nasjonale strategi med i betraktningen ved planlegging og gjennomføringen av denne studien. EU-kommisjonen har nedsatt en ekspertgruppe som har utarbeidet etiske retningslinjer for pålitelig bruk av KI. De har foreslått syv prinsipper for etisk og ansvarlig utvikling av KI (AI, 2019):

- KI-baserte løsninger skal respektere menneskets selvbestemmelse og kontroll
- Mennesker skal være inne i beslutningsprosessene for å kvalitetssikre og gi tilbakemelding i alle ledd i prosessen
- KI-baserte systemer skal være sikre og teknisk robuste
- KI skal ta hensyn til personvernet
- KI-baserte systemer må være gjennomsiktige
- KI-systemer skal legge til rette for inklusjon, mangfold og likebehandling
- KI skal være nyttig for samfunn og miljø
- Ansvarlighet

Regjeringen trakk i den nasjonale strategien frem at KI skal bygge på etiske prinsipper, respektere menneskerettighetene og ivareta den enkeltes personvern. Dette kan potensielt medføre utfordringer ved bruk av KI til forskning i det medisinske fagfeltet, og det kan på enkelte områder være behov for å utvikle regelverk før man prøver ut metoder basert på KI. Forskning gir i mange tilfeller programvarene tilgang til store mengder data, noe som er særlig utfordrende dersom dataene inneholder personopplysninger. Fordelen med systematiske oversiktsartikler er at de ofte ikke inneholder direkte personopplysninger, da opplysningene som benyttes allerede er publisert. Dette gjelder også for denne oppgaven. Bruk av anonymiserte data trekkes i den nasjonale strategien frem som et mer personvernvennlig alternativ. I tillegg var dette en metodestudie, hvor resultatene ikke gir direkte implikasjoner på kliniske anbefalinger og pasientbehandling. Regjeringen presiserte også at de ønsker at vi i Norge forsker på KI. Gjennomføring av denne oppgaven ble derfor vurdert til å være i tråd med den nasjonale strategien og til å ivareta etiske prinsipper og personvern.

Da det til denne oppgaven ikke skal innhentes personopplysninger kreves det ingen søknad til eksterne godkjenningsinstanser, som Norsk senter for forskningsdata (NSD) eller Regionale komiteer for medisinsk og helsefaglig forskningsetikk (REK). Protokollen for gjennomføring av den systematiske oversikten er allerede publisert og tilgjengelig online, noe som ivaretar prinsippet om åpenhet i forskning (Forskningsrådet, 2020).

5. Resultat

5.1. Stjernerangeringer

Etter gjennomføring av det systematiske litteratursøket identifiserte forfatterne av den systematiske oversiktsartikkelen 7994 treff som alle er inkludert i min oppgave. Etter menneskelig vurdering av samtlige artikler ble 127 inkludert for lesing i fulltekst. Disse er alle kategorisert som relevante for inklusjon i denne oppgaven. Dette tilsvarer en punktprevalens på 1,6%. Rayyan kalkulerte stjernerangeringer basert på de menneskelige vurderingene for hvert 20% nivå (Tabell 4).

Tabell 4. Antall artikler per stjernerangeringer kalkulert av Rayyan etter manuell sortering av artikler i 20% nivåer.

Artikler annotert	Stjernerangering				
	★★★★★	★★★★★	★★★★★	★★★★★	★★★★★
1609 (20%)	1716 (21%)	677 (8%)	3991 (50%)	1 (0%)	0 (0%)
3198 (40%)	1554 (19%)	1078 (13%)	2146 (27%)	18 (0%)	0 (0%)
4821 (60%)	1313 (16%)	749 (9%)	1082 (14%)	25 (0%)	4 (0%)
6400 (80%)	738 (9%)	335 (4%)	507 (6%)	12 (0%)	2 (0%)

5.2. Prestasjonen til maskinlæringsverktøyet

De samlede resultatene for sensitivitet, spesifisitet, PPV, NPV og PR for alle 20% nivåene er presentert i tabell 5. Denne tabellen gir også mulighet for sammenligning av de to grenseverdiene som ble benyttet til denne oppgaven. I de neste avsnittene vil jeg gi en detaljert presentasjon av resultatene for hvert 20% nivå.

Tabell 5. Resultater for sensitivitet, spesifisitet, positiv prediktiv verdi (PPV), negativ prediktiv verdi (NPV) og punktprevalens for alle 20% nivåer.

Ekskludert		**Manuelt sortert**			
		20%	**40%**	**60%**	**80%**
Sensitivitet		96.2%	98.7%	100%	100%
		0%	18.2%	54.6%	52.6%
Spesifisitet		38%	55.8%	66.9%	68.1%
		100%	99.9%	99.8%	99.8%
PPV		2.5%	3.5%	4%	3.6%
		0%	77.8%	82.8%	71.4%
NPV		99.8%	100%	100%	100%
		98.4%	98.7%	99.4%	99.4%
Punktprevalens		1.6%	1.6%	1.4%	1.2%
		1.6%	1.6%	1.4%	1.2%

5.3. 20% av studiene manuelt vurdert

5.3.1. Grenseverdi <2.5 og ≤ 2.5 stjerner for eksklusjon

Etter manuell sortering av 20% av artiklene var 1609 artikler sortert til kategoriene ekskludert eller inkludert (Tabell 4). Av disse ble 1587 ekskludert, mens 22 artikler ble inkludert (17.3% av de inkluderte artiklene). Maskinlæringsverktøyet i Rayyan beregnet deretter stjernerangeringene for de resterende 6385 artiklene. Dette resulterte i at 2393 (29%) artikler fikk <2.5 stjerner og 6384 (79%) artikler fikk ≤2.5 stjerner (Tabell 4). Ved en grenseverdi på <2.5 stjerner for eksklusjon oppnår maskinlæringsverktøyet sensitivitet på 96.2%, spesifisitet på 38%, PPV på 2.5% og NPV på 99.8% (Tabell 2). Dersom det heller ble benyttet en grenseverdi på ≤2.5 stjerner for eksklusjon ble det oppnådd sensitivitet på 0%, spesifisitet på 100%, PPV på 0% og NPV på 98.4% (Tabell 2). Det vil si at Rayyan ikke evnet å identifisere alle relevante artikler ved noen av grenseverdiene etter manuell sortering av 20%. NPV er derimot høy ved begge grenseverdiene, og spesifisiteten med en grenseverdi på ≤2.5 stjerner for eksklusjon.

5.4. 40% av studiene manuelt vurdert

5.4.1. Grenseverdi <2.5 og ≤ 2.5 stjerner for eksklusjon

Etter manuell sortering av 40% av artiklene var 3198 artikler sortert. Av disse ble 3148 ekskludert, mens 50 artikler ble inkludert (39.4% av de inkluderte artiklene). Etter generering av stjernevurderingene fikk 2632 (22%) artikler <2.5 stjerner og 4778 (59%) artikler ≤2.5 stjerner (Tabell 4). Rayyan oppnådde da med en grenseverdi på <2.5 stjerner for eksklusjon en sensitivitet 98.7%, spesifisitet på 55.8%, PPV på 3.5% og NPV på 100%. Dersom det heller ble benyttet en grenseverdi på ≤ 2.5 stjerner for eksklusjon var sensitiviteten 18.2%, spesifisiteten 99.9%, PPV 77.8% og NPV 98.7%. Sensitiviteten økte dermed noe etter sortering av 40% av artiklene, men fremdeles evnet ikke maskinlæringsverktøyet å identifisere alle relevante artikler ved noen av disse grenseverdiene. Med en grenseverdi på ≤2.5 for eksklusjon ses også en økning i PPV som betyr at evnen til å identifisere relevante artikler var forbedret sammenlignet med etter vurdering av 20%.

5.5. 60% av studiene manuelt vurdert

5.5.1. Grenseverdi <2.5 og ≤ 2.5 stjerner for eksklusjon

Etter manuell sortering av 60% av artiklene var 4821 artikler sortert. Av disse ble 4738 ekskludert, mens 83 artikler ble inkludert (65.4% av de inkluderte artiklene). Rayyan beregnet deretter stjernerangeringene som resulterte i at 2062 (25%) artikler fikk >2.5 stjerner og 3144 (39%) artikler fikk ≤2.5 stjerner (Tabell 4). Maskinlæringsverktøyet oppnådde med en grenseverdi på <2.5 stjerner for eksklusjon en sensitivitetsverdi på 100%, spesifisitet på 65.9%, PPV på 4% og NPV på 100% (Tabell 5). Med en stjerneverdi på ≤ 2.5 stjerner for eksklusjon var sensitiviteten 54.6%, spesifisiteten 99.8%, PPV 82.8% og NPV 99.4%. Med en grenseverdi på <2.5 stjerner for eksklusjon ble altså alle relevante artikler identifisert. Dette er også det nivået av manuell sortering som førte til best spesifisitet ved denne grenseverdien.

5.6. 80% av studiene manuelt vurdert

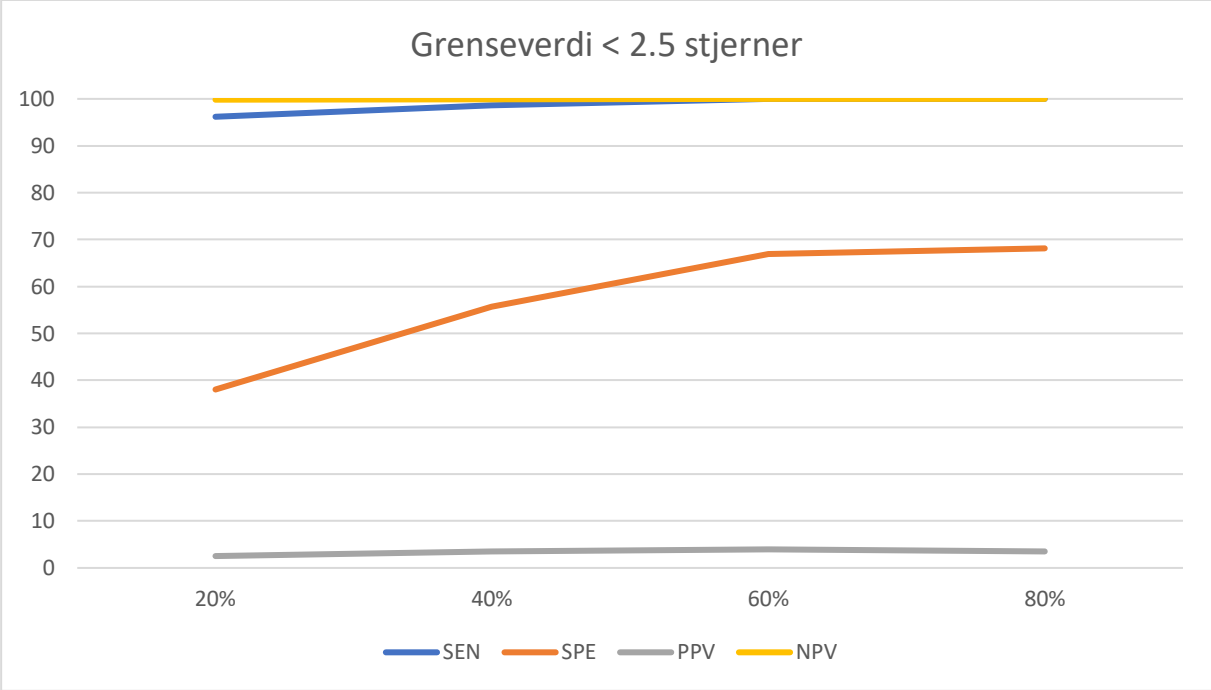
5.6.1. Grenseverdi <2.5 og ≤ 2.5 stjerner for eksklusjon

Etter manuell sortering av 80% av artiklene var 6400 artikler sortert. Av disse ble 6292 ekskludert, mens 108 ble inkludert (85% av de inkluderte artiklene). Dette resulterte i at 1073 (13%) artikler fikk >2.5 stjerner og 1580 (19%) artikler fikk ≤ 2.5 stjerner (Tabell 4). Rayyan oppnådde med en grenseverdi på <2.5 stjerner for eksklusjon dermed sensitivitet på 100%, spesifisitet på 68.1%, PPV på 3.7% og NPV på 100% (Tabell 5). Med en grenseverdi på ≤ 2.5 stjerner for eksklusjon var sensitiviteten 52.6%, spesifisiteten 99.8%, PPV 71.4% og NPV 99.4%. Det var dermed minimale endringer i utfallsmålene etter sortering av 80% av artiklene sammenlignet med etter sortering av 60%. Enkelte verdier viste noe nedgang, som vil si at maskinlæringsverktøyets evne til å ekskludere irrelevante artikler ikke ble bedre av å manuelt sortere flere artikler på dette tidspunktet.

5.7. Oppsummering

5.7.1. Grenseverdi <2.5 stjerner for eksklusjon

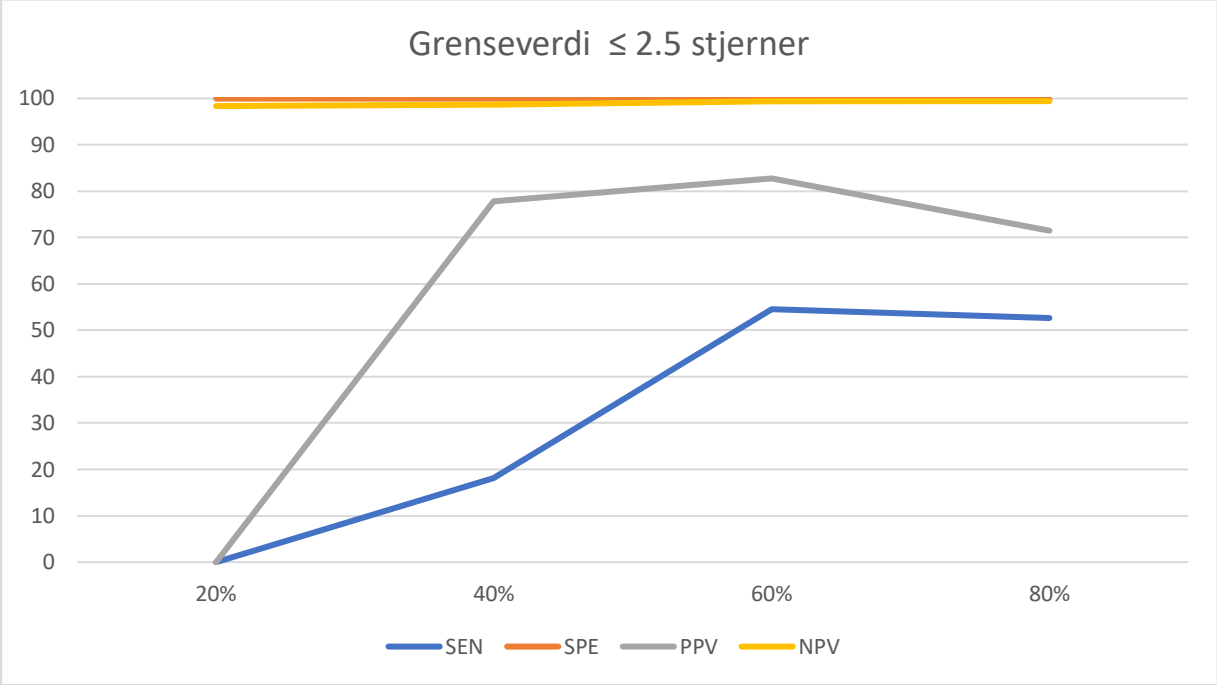
Samlede resultater for en grenseverdi på <2.5 stjerner for eksklusjon er presentert i figur 5. Med denne grenseverdien oppnådde Rayyan svært god sensitivitet, mens spesifisiteten maksimalt nådde 68.1% etter at 80% av studiene var manuelt sortert. Tilsvarende kan ses ved sammenligning av PPV og NPV: Rayyan oppnådde lav PPV på alle stadiene av manuell sortering. NPV verdiene var derimot svært høye. Disse resultatene viste at Rayyan med en grenseverdi på <2.5 hadde god evne til å ekskludere studier som ikke var relevante for inklusjon, men relativt dårlig evne til å identifisere studiene som skulle inkluderes. Som illustrert i figuren skjer det svært lite endring i resultatene mellom nivåene hvor 60% og 80% av artiklene var manuelt sortert.



Figur 5. Samlede resultater med grenseverdi <2.5 stjerner for eksklusjon.

5.7.1. Grenseverdi ≤2.5 stjerner for eksklusjon

De samlede resultatene for en grenseverdi ≤ 2.5 stjerner for eksklusjon er presentert i figur 6. Med en grenseverdi på ≤ 2.5 stjerner for eksklusjon hadde Rayyan nær perfekt spesifisitet, men oppnådde maksimalt 54.6% sensitivitet. Også NPV var svært høy på alle stadiene av den manuelle sorteringen. PPV var ved denne grenseverdien 0% etter at 20% av artiklene var sortert og steg til maksimalt 82.8%. Dette etter at 60% av studiene var manuelt sortert. Også ved denne grenseverdiene ses det liten forskjell mellom de to øverste nivåene. Det må også bemerkes at lav sensitivitet ved alle stadier av manuell vurdering for denne grenseverdien tyder på at en betydelig andel relevante studier ville blitt feilaktig ekskludert på alle nivåer.



Figur 6. Samlede resultater med grenseverdi ≤ 2.5 stjerner for eksklusjon.

6. Diskusjon

6.1. Oppsummering av resultatene

Det implementerte maskinlæringsverktøyet i Rayyan viste god prestasjon når det kom til å ekskludere irrelevante artikler ved en grenseverdi på <2.5 for eksklusjon, men evnen til å identifisere relevante artikler var dårlig. Med en grenseverdi på ≤2.5 var spesifisiteten bedre, men en betydelig andel relevante artikler ville blitt feilaktig ekskludert ved automatisering av oppgaven. Resultatene ved de to grenseverdiene varierte altså i stor grad og hvis maskinlæringsverktøyet skal benyttes ved studieseleksjon til systematiske oversikter er valg av grenseverdi avgjørende for resultatet. For studieseleksjon etter et litteratursøk til en systematisk oversiktsartikkel som omhandler prognostiske modeller indikerer resultatene i denne oppgaven at en grenseverdi på <2.5 kan være aktuell å benytte, og vil medføre en lav risiko for å miste relevante artikler.

6.2. Tolkning av resultatene – valg av grenseverdier

6.2.1. Andre grenseverdier?

De to grenseverdiene ble valgt a priori på bakgrunn av tidligere forskning som har vist at en grenseverdi på <2.5 stjerner for eksklusjon maksimerer sensitiviteten (Li et al., 2023; Valizadeh et al., 2022). Det kan også antas at artiklene med denne verdien er vanskeligst for maskinlæringsverktøyet å vurdere. Som det fremkom av resultatene var det stor forskjell i sensitivitet og spesifisitet avhengig av hvilken av disse grenseverdiene som ble benyttet. Det ville derfor vært lite relevant å undersøke utfallsmålene ved andre grenseverdier for inklusjon og eksklusjon. En høyere grenseverdi ville ført til ytterligere fall i sensitivitet, og en lavere grenseverdi ville senket spesifisiteten uten en betydningsfull endring i sensitiviteten. Ved en høyere grenseverdi (≤3.5 stjerner for eksklusjon) identifiserte maskinlæringsverktøyet maksimalt 4 artikler som relevante for inklusjon (etter manuell sortering av 60%) (Tabell 4). Det vil derfor ikke være aktuelt å benytte verktøyet ved en høyere grenseverdi. Mens ved en lavere grenseverdi for eksklusjon vil man måtte lese gjennom en større andel irrelevante studier uten å identifisere flere relevante, og derfor ikke oppnå en optimal reduksjon i arbeidsbelastningen. Disse resultatene samsvarer med resultatene fra Valizadeh et al (2022) noe som styrker denne beslutningen.

6.2.2. Sensitivitet og spesifisitet

6.2.2.1. Grenseverdi <2.5

Når Rayyan ble benyttet med en grenseverdi på <2.5 for eksklusjon, hvor artikler med en stjernerangering på 2.5 eller høyere ble vurdert som relevante, var sensitiviteten 100%, mens spesifisiteten maksimalt nådde 68.1%. Fordelen med perfekt sensitivitet er at alle relevante studier identifiseres dersom artiklene med 2.5 stjerner eller mer vurderes for inklusjon. Høy sensitivitet bør derfor prioriteres fremfor høy spesifisitet ved seleksjon av artikler til en systematisk oversikt. Det er likevel viktig å være klar over at den høye verdien for sensitivitet førte til en kraftig reduksjon i spesifisitet. Det er også verdt å merke seg at sensitivitet på 100% ble oppnådd først etter at 60% av studiene var manuelt sortert. Spesifisiteten var da 66.9%. Det vil si at etter menneskelig vurdering av 4821 artikler gjenstod det fremdeles 1111 sammendrag som måtte blitt lest og vurdert ved bruk av tradisjonelle metoder dersom artiklene med >2.5 stjerner ble inkludert. Dette er en tidkrevende prosess og det er derfor aktuelt å spørre seg hvor tidsbesparende det er å benytte maskinlæringsverktøyet ved denne grenseverdien dersom 100% sensitivitet er nødvendig. Allerede etter manuell sortering av 20% av artiklene oppnådde maskinlæringsverktøyet en sensitivitetsverdi på 96.2%. For å redusere arbeidsbelastningen ytterligere er det derfor viktig å diskutere hvor stor reduksjon i sensitivitet som eventuelt kan aksepteres, eller benytte andre metoder for å forbedre prestasjonen til maskinlæringsverktøyet og dermed redusere arbeidsbelastningen ytterligere. Mulighetene for dette diskuteres videre i kapittel 6.4.2.

6.2.2.2. Grenseverdi ≤ 2.5

Ved en grenseverdi på ≤2.5 for eksklusjon, hvor studier med en stjernerangering på 2.5 ble vurdert som irrelevante, oppnådde Rayyan maksimalt 54.6% sensitivitet. Spesifisiteten var da 99.8%. Disse verdiene ble oppnådd etter manuell sortering av 60% av artiklene. Det vil si at 20 av artiklene som etter menneskelige vurderinger ble inkludert i den systematiske oversiktsartikkelen ikke ville blitt identifisert ved bruk av maskinlæringsverktøyet i Rayyan. Maskinlæringsverktøyet hadde ved denne grenseverdien dermed god evne til å ekskludere studier som ikke var relevante for inklusjon, men relativt dårlig evne til å identifisere studiene som skulle inkluderes. Det er da viktig å vurdere hvor stor betydning artiklene man potensielt mister har ved utarbeiding av systematiske oversiktsartikler, da dette potensielt kan få store

konsekvenser for de endelige resultatene. For å sikre tilfredsstillende validitet og reliabilitet burde alle relevante artikler identifiseres (Shemilt et al., 2016). Derfor burde grenseverdien som maksimerer sensitiviteten benyttes selv om dette går på bekostning av spesifisiteten. Til bruk for utarbeiding av systematiske oversiktsartikler vil jeg basert på disse resultatene argumentere for at maskinlæringsverktøyet i Rayyan ikke kan brukes til å automatisk ekskludere artikler med en stjernerangering på ≤2.5 stjerner, da sensitivitetsverdien er for lav og risikoen for feil dermed for stor til å aksepteres.

Tidligere studier har vist at de viktigste studiene ofte får en høy stjernerangering (Olofsson et al., 2017). Et alternativ kan derfor være å benytte Rayyan med en grenseverdi på ≤2.5 når det ikke er like essensielt å identifisere alle relevante studier, som for eksempel i situasjoner hvor man raskt ønsker å få et overblikk over den viktigste litteraturen innen et emne. Dette kan være aktuelt for å identifisere kunnskapshull når behovet for utarbeiding av en ny systematisk oversiktsartikkel skal undersøkes eller ved utarbeiding av en litteraturgjennomgang eller en sonderende oversiktsartikkel (scoping review). Ved utarbeiding av sonderende oversiktsartikler kartlegges det eksisterende kunnskapsgrunnlaget ved at man henter inn kunnskap med større vekt på bredere og mindre spesifikke søk sammenlignet med ved utarbeiding av systematiske oversiktsartikler (Munn et al., 2018). Det kan derfor være hensiktsmessig å da benytte en høyere grenseverdi for å raskt identifisere de mest relevante artiklene.

Oppsummert viste resultatene for sensitivitet og spesifisitet fra denne oppgaven at maskinlæringsverktøyet i Rayyan potensielt kan benyttes for studieseleksjon til en systematisk oversiktsartikkel på prognostiske modeller for degenerativ ryggkirurgi. Grenseverdien på <2.5 stjerner for eksklusjon er da mest aktuell å benytte da dette maksimerer sensitivitetsverdien. En høyere grenseverdi kan eventuelt benyttes i situasjoner hvor det ikke er like essensielt å identifisere alle relevante artikler, men kun de viktigste. Ved begge grenseverdiene er det uansett behov for menneskelige vurderinger for å øke spesifisiteten, og den potensielle reduksjonen i arbeidsbelastningen er fremdeles usikker.

6.2.3. PPV og NPV

Med en grenseverdi på <2.5 stjerner for eksklusjon oppnådde maskinlæringsalgoritmen i Rayyan svært høye NPV verdier, mens PPV var lave på alle nivåene av manuell sortering. Dersom grenseverdien på ≤ 2.5 stjerner for eksklusjon derimot ble benyttet oppnådde maskinlæringsalgoritmen maksimalt PPV på 82.8%, og dette først etter at 60% av studiene var manuelt sortert. NPV er på den andre siden svært høy på alle nivåene av manuell sortering. Det er viktig å merke seg at PPV og NPV i stor grad er avhengig av prevalensen av artikler som er aktuelle for inklusjon, som igjen avhenger av sensitiviteten til søkestrategien. Derfor ble også punktprevalensen for hvert nivå kalkulert. Det er utfordrende å utforme en presis søkestrategi for prognostiske studier, noe som medførte en lav punktprevalens på alle nivåer. Dette er en vanlig problemstilling ved utarbeiding av systematiske oversiktsartikler i det medisinske fagfeltet, og vanligvis vil omtrent 97% av artiklene være irrelevante (Borah et al., 2017). Den lave prevalensen av relevante artikler var sannsynligvis årsaken til høy NPV ved alle nivåer av manuell sortering uavhengig av grenseverdi. Ved bruk av veiledet læring på så skjevfordelte data risikerer man også at maskinlæringsverktøyet feilaktig ekskluderer relevante studier for å bedre prestasjonen, da dette er det mest sannsynlige utfallet (van Dinter et al., 2021). Med en så lav punktprevalens kan det stilles spørsmål ved hvor betydningsfulle disse verdiene egentlig er for å vurdere prestasjonen til maskinlæringsverktøyet, og det vil videre i diskusjonen derfor legges mer vekt på sensitivitets- og spesifisitetsverdiene.

6.3. Sammenligning med tidligere studier

6.3.1. Tidligere studier på Rayyan

Flere tidligere studier har undersøkt prestasjonen til maskinlæringsverktøyet i Rayyan ved bruk av tilsvarende utfallsmål som i denne oppgaven, og er derfor aktuelle å sammenligne med (Tabell 2). Studien gjennomført av Valizadeh et al (2022) er den som metodisk har flest fellestrekk med metoden som ble benyttet til denne oppgaven, da de benyttet lignende fremgangsmåte og samme utfallsmål. I tillegg benyttet de søkeresultatene fra tre systematiske oversiktsartikler på diagnostiske tester, som har fellestrekk med studier på prognostiske modeller da de ikke tildeles emneord i søkedatabaser og ofte har lav punktprevalens. Med en grenseverdi på <2.5 stjerner for eksklusjon oppnådde Rayyan i deres studie sensitivitetsverdier på 97-99% med spesifisitetsverdier på 19-58%. Dette samsvarer i

stor grad med resultatene fra denne oppgaven både for sensitivitet (96%-100) og spesifisitet (38%-68.1%). Tilsvarende fant de også at en grenseverdi på ≤2.5 stjerner resulterte i perfekt spesifisitet men lav sensitivitet, og forfatterne konkluderte med at Rayyan er et reliabelt verktøy for å ekskludere irrelevante artikler men presterer dårlig for identifisering av relevante artikler (Valizadeh et al., 2022). Den store forskjellen på resultatene avhengig av hvilken grenseverdi som benyttes indikerer at programvaren har dårlig evne til å kategorisere artiklene rundt denne grenseverdien. Både i den oppgaven og i den nevnte studien ble prestasjonen til Rayyan vurdert ved studieseleksjon til systematiske oversiktsartikler innen det medisinske fagfeltet, og de inkluderte artiklene hadde som nevnt flere likhetstrekk. Likevel ble studiene gjennomført på ulike populasjoner. Det er derfor et interessant funn at Rayyan presterer relativt likt uavhengig av populasjon ved seleksjon av denne typen artikler.

Samtidig skal det nevnes at resultatene for PPV og NPV var vesentlig forskjellig fra hva som ble funnet i denne oppgaven. Med en grenseverdi på <2.5 stjerner for eksklusjon fant Valizadeh (2022) en PPV på 22-31% og NPV på 98-100%. Tilsvarende resultater fra denne oppgaven viste en PPV på 2.5-4% og NPV på 99.8-100%. Dersom man heller sammenligner resultatene ved en grenseverdi på ≤2.5 stjerner for eksklusjon fant Valizadeh (2022) en PPV på 86-97% og NPV på 56-88%, mens resultatene fra denne oppgaven viste en PPV på 0-82.8% og NPV på 98.4-99.4%. PPV og NPV har som nevnt i stor grad sammenheng med presisjonen til det systematiske litteratursøket. Begge disse studiene ble gjennomført på systematiske oversiktsartikler hvor det typisk er lav presisjon og dermed lav punktprevalens. Det var likevel en vesentlig forskjell i punktprevalensen mellom de to studiene. I denne oppgaven varierte punktprevalensen ved de ulike nivåene fra 1.2-1.6%. Til sammenligning hadde litteratursøkene som ble brukt av Valizadeh (2022) en punktprevalens på 13-22%, som er en høyere prevalens av relevante artikler enn det som vanligvis oppnås ved litteratursøk til systematiske oversiktsartikler (Li et al., 2023). Dette er trolig den viktigste årsaken til forskjellen i NPV og PPV.

Li et al (2023) rapporterte også sensitivitets- og spesifisitetsverdiene ved en grenseverdi på <2.5 stjerner for eksklusjon etter menneskelig vurdering av ca. 30% av artiklene til en systematisk oversiktsartikkel. I deres studie oppnådde maskinlæringsverktøyet i Rayyan en sensitivitet på 97% og en spesifisitet på 95.3%

ved en grenseverdi på <2.5 stjerner for eksklusjon. Spesifisiteten var dermed betydelig høyere enn det det som ble oppnådd i denne oppgaven, og den totale prestasjonen til maskinlæringsverktøyet svært god. De benyttet et datasett på 500 artikler, hvorav 10% var relevante for inklusjon. De trakk frem at mer presise estimater av sensitivitet og spesifisitet potensielt kan oppnås med større datasett (Li et al., 2023). Olofsson et al (2017) brukte seks datasett som inneholdt totalt 7956 artikler til sine analyser. De oppga ikke hvilken grenseverdi som ble benyttet for eksklusjon, men rapporterte sensitivitets- og spesifisitetsverdier etter menneskelig vurdering av 25%, 50% og 75% av artiklene. De fant at medianverdien for sensitivitet var 60% etter at 25% av artiklene var manuelt vurdert og 95% etter at halvparten av artiklene var manuelt vurdert. Etter at 25% var manuelt vurdert varierte sensitiviteten fra 21% til 99%. Resultatene så ut til å variere avhengig av presisjonen til litteratursøket og størrelse på studien. Basert på dette, resultatene fra min oppgave og fra studien til Li (2023) ser det ut til at størrelsen på datasettet, punktprevalensen og seleksjonskriteriene potensielt kan påvirke prestasjonen til Rayyan. Dette gjelder særlig tidlig i seleksjonsprosessen.

Flere av de andre studiene som er gjort på Rayyan fokuserer på brukernes opplevelse av programvaren og tidsbesparelse heller enn på prestasjonen til maskinlæringsalgoritmen. Resultatene kan derfor ikke direkte sammenlignes med resultatene fra denne oppgaven, men kan likevel være et nyttig supplement ved diskusjon om Rayyan bør implementeres ved utarbeiding av systematiske oversiktsartikler. Utviklerne av programvaren konkluderte i sin pilotstudie med at programvaren er intuitiv å bruke og kan redusere tidsbruken med over 50% (Ouzzani et al., 2016). Andre fordeler som ble trukket frem med Rayyan er at programvaren har mange tilleggsfunksjoner, presterer godt på brukerundersøkelser, ikke krever kunnskap om programmering og er enkel å bruke sammenlignet med flere av de andre alternativene (Harrison et al., 2020). Dette var også hovedargumentene for at programvaren ble valgt til denne oppgaven. Det at Rayyan er gratis og tilgjengelig online er andre gode argumenter for å velge denne programvaren over andre programvarer med integrerte maskinlæringsverktøy. Det er da viktig å undersøke og sammenligne prestasjonen til de aktuelle maskinlæringsverktøyene.

6.3.2. Tidligere studier på andre maskinlæringsverktøy

Det kan være vanskelig å orientere seg i det store utvalget av ulike maskinlæringsverktøy som finnes og mulighetene de tilfører. Av andre populære alternativer til Rayyan er det særlig Abstrackr, EPPI-reviewer, DistillerSR, RobotAnalyst og Research Screener som trekkes frem i tidligere forskning (Cierco Jimenez et al., 2022; Valizadeh et al., 2022). Resultatene fra tidligere studier på disse programvarene kan brukes til å sammenligne prestasjonen til de ulike maskinlæringsverktøyene med prestasjonen til maskinlæringsverktøyet i Rayyan, og bidra til økt kunnskap om hvilke som er mest aktuelle å benytte ved studieseleksjon til systematiske oversiktsartikler.

Det ble i 2018 publisert en artikkel som rapporterte sensitivitets- og spesifisitetsverdier for maskinlæringsverktøyet i Abstrackr (Gates et al.). Resultatene fra denne studien viste sensitivitetsverdier på 79-92% og spesifisitetsverdier på 69-90% etter menneskelig vurdering av 0.7-10.3% av artiklene. Sammenlignet med resultatene fra denne oppgaven oppnådde Abstrackr da bedre samlede sensitivitets- og spesifisitetsverdier etter at en tilsvarende andel av artiklene var sortert. Også i artikkelen fra Li et al (2023) oppnådde Abstrackr bedre spesifisitetsverdier enn Rayyan, selv om begge verktøyene potensielt kan redusere arbeidsbelastningen og bidra til hurtigere ferdigstilling av systematiske oversiktsartikler. En annen studie på prestasjonen til Abstrackr viste derimot at en større andel av artiklene måtte vurderes for å oppnå 100% sensitivitet sammenlignet med ved bruk av Rayyan (Tsou et al., 2020). Gates (2019) har senere publisert en annen artikkel hvor prestasjonen til Abstrackr ble sammenlignet med RobotAnalyst og DistillerSR. I denne artikkelen oppnådde Abstrackr en gjennomsnittlig sensitivitet på 89% etter manuell vurdering av 1.7-3.4% av artiklene, og presterte betydelig bedre enn de to andre verktøyene. Abstrackr presterte også her bedre enn Rayyan gjorde i denne oppgaven, men spesifisitetsverdier var ikke rapportert. Harrison et al (2020) anbefalte bruk av Rayyan fremfor Abstrackr for studieseleksjon til systematiske oversiktsartikler basert på sin brukerundersøkelse. Samlet sett ser det derfor ut til at både Abstrackr og Rayyan potensielt kan benyttes, men det er usikkert hvilken av programvarene som presterer best. DistillerSR og RobotAnalyst ser derimot ut til å prestere dårligere både når det kommer til prestasjon og brukervennlighet (Cleo et al., 2019; Przybyła et al., 2018).

En annen artikkel har sammenlignet prestasjonen til Abstrackr med maskinlæringsverktøyet EPPI-Reviewer (Tsou et al., 2020). Forfatterne konkluderte, etter å ha testet verktøyene på ni ulike oversiktsartikler, med at begge presterte godt. EPPI-Reviewer viste likevel jevnt over best resultater. Ved bruk av EPPI-Reviewer oppnådde de en sensitivitet på 100% etter manuell vurdering av 39.9-98.8% av artiklene, med en gjennomsnittlig verdi på 83.4%. Det var dermed relativt stor variasjon i prestasjonen avhengig av hvilken systematisk oversiktsartikkel maskinlæringsverktøyet ble testet på. Til sammenligning oppnådde jeg i denne oppgaven samme sensitivitetsverdi etter manuell vurdering av 60% med en stjerneverdi på <2.5, mens sensitiviteten maksimalt nådde 54.6% med en stjerneverdi på $\leq$2.5 for eksklusjon. EPPI-Reviewer viste dermed også lovende resultater og kan benyttes for å redusere arbeidsbelastningen ved utarbeiding av systematiske oversiktsartikler. Likevel presterte både EPPI-Reviewer og Abstrackr dårligere enn Rayyan på brukerundersøkelsen gjennomført av Harrison et al (2020). EPPI-Reviewer kan dermed potensielt være et bedre alternativ sammenlignet med Abstrackr, mens det er behov for mer forskning for å sammenligne prestasjonen til EPPI-Reviewer og Rayyan på de samme datasettene.

Research screener er et nyere maskinlæringsverktøy som bruker dyplæring, naturlig språkbehandling og andre maskinlæringsmetoder for å semi-automatisere studieseleksjonsprosessen ved utarbeiding av systematiske oversiktsartikler. Dyplæring er en mer avansert form for KI sammenlignet med naturlig språkbehandling og veiledet læring som benyttes i Rayyan, og kan derfor potensielt medføre flere muligheter og mer presise estimater. Forskerne som utviklet verktøyet testet det på ni ulike systematiske litteraturstudier, og fant at Research Screener var i stand til å identifisere relevante sammendrag med høy grad av nøyaktighet. I gjennomsnitt måtte 20.7% av artiklene manuelt sorteres før alle relevante artikler var identifisert (Chai et al., 2021). Research Screener presterte dermed betydelig bedre enn maskinlæringsverktøyet til Rayyan gjorde i denne oppgaven, hvor 60% av artiklene måtte manuelt sorteres før sensitiviteten nådde 100%. Dette er ikke overaskende da det er kjent at støttevektormaskiner, som benyttes i Rayyan, ikke er ansett som optimale klassifiseringsalgoritmer til dette formålet (Valizadeh et al., 2022). Det er derfor interessant at den mer avanserte teknologien i Research Screener tilsynelatende førte til bedre prestasjon. Det er likevel verdt å merke seg at

den eneste identifiserte studien som undersøker prestasjonen til Research screener er publisert av utviklerne av programvaren. Det er derfor behov for videre forskning på prestasjonen til Research Screener gjennomført av uavhengige forskere for å redusere risikoen for at resultatene kan være påvirket av en interessekonflikt.

Basert på disse studiene er det vanskelig å konkludere med hvilken programvare som burde benyttes ved studieseleksjon til systematiske oversiktsartikler, da det ikke er en tydelig konsensus vedrørende prestasjonen til de ulike maskinlæringsverktøyene. Rayyan trekkes i flere studier frem som et godt alternativ, særlig blant de kostnadsfrie alternativene, men også flere av de andre maskinlæringsvertøyene har prestert godt. Særlig Research Screener som benytter mer avanserte former for KI har vist lovende resultater (Chai et al., 2021). Likevel viste resultatene både fra denne oppgaven og samtlige tidligere studier på prestasjonen til Rayyan og andre maskinlæringsverktøy at disse potensielt kan redusere arbeidsbelastningen og bidra til hurtigere produksjon av systematiske oversiktsartikler.

6..4. Stoppkriterier og redusert arbeidsbelastning

Det er flere metoder som kan benyttes for å optimalisere prestasjonen til maskinlæringsverktøyene. Et av disse er stoppkriterier, som vil si at studieseleksjonen avsluttes før alle treff i litteratursøket er menneskelig vurdert. Olofsson et al (2017) diskuterte i sin artikkel implementering av spesifikke kriterier for å avslutte studieseleksjonen basert på stjernerangeringene ved bruk av Rayyan. De konkluderte med at det fremdeles er behov for mer forskning før eventuelle stoppkriterier kan tas i bruk på grunnlag av et lite utvalg i deres studie, behov for valide metoder ved utarbeiding av systematiske oversiktsartikler og målet om å identifisere alle relevante artikler. For å utnytte mulighetene programvarene fører med seg i enda større grad er det likevel nødvendig å enes om stoppkriterier for å muliggjøre automatisering på nivå 3 og 4 (Figur 2) (O'Connor et al., 2019). Disse stoppkriteriene kan sammen med andre metoder for å optimalisere prestasjonen til maskinlæringsverktøyene potensielt bidra til å ytterligere redusere arbeidsbelastningen ved studieseleksjon til systematiske oversiktsartikler.

6.4.1. Hvor stor andel av studiene må vurderes manuelt?

For at maskinlæringsverktøyet i Rayyan skal ha mulighet til å kalkulere stjernerangeringer er det avhengig av å lære av menneskelige vurderinger. Jo flere artikler som blir vurdert, jo mer lærer algoritmen. Derfor kan man anta at kalkuleringene vil bli mer presise dersom flere artikler blir vurdert. Dette medfører samtidig en større arbeidsbelastning da gjennomføring av de menneskelige vurderingene er tidkrevende. Ifølge Cochrane håndboken skal det ved utarbeiding av systematiske oversiktsartikler benyttes metoder som maksimerer sensitiviteten (Lefebvre et al., 2019). Tidligere studier har foreslått at sensitivitet på 99% kan være akseptabelt ved utarbeiding av systematiske oversiktsartikler, men selv feilaktig eksklusjon av en relevant artikkel kan føre til en betydningsfull endring i resultatene (Thomas et al., 2017). Dersom stoppriterier skal benyttes er det derfor ønskelig å identifisere den optimale andelen studier som må menneskelig vurderes for å oppnå akseptable sensitivitetsverdier, og metoder som reduserer sensitiviteten bør vurderes kritisk. Det finnes flere alternative stoppkriterier som kan benyttes for å bestemme når seleksjonen av artikler studier skal avsluttes. En artikkel fra 2020 legger frem forslag til stoppkriterier fra fire kategorier (Callaghan & Müller-Hansen):

- Forhåndsbestemte utvalgskriterier: Forfatterne estimerer hvor mange artikler de forventer å identifisere, og stopper når dette antallet er nådd (Shemilt et al., 2014).
- Heuristisk: Vurderingene stopper etter identifisering av et gitt antall påfølgende irrelevante artikler (Jonnalagadda & Petitti, 2013)
- Pragmatiske kriterier: Vurderingen av artikler avsluttes når forfatterne ikke har mer tid til å gjennomføre dette (Miwa et al., 2014)
- Automatiske stoppkriterier: Mer kompliserte automatiserte systemer avgjør når vurderingene skal avsluttes (Di Nunzio, 2018; Yu & Menzies, 2019)

Tidligere studier har sett at det å vurdere en viss prosentandel av de relevante artiklene, for eksempel 25% nivåer, fører til mest reliabel prestasjon (Scherhag & Burgard, 2023). Dette samsvarer med fremgangsmetoden som ble brukt i denne oppgaven, og det kan på bakgrunn av dette antas at prestasjonen er reliabel selv om andre metoder også kunne blitt benyttet.

Det er ved bruk av Research Screener vist at alle relevante artikler var identifisert etter manuell sortering av 4-40% av artiklene. Forfatterne foreslo på bakgrunn av dette en konservativ grense hvor 50% vurderes manuelt for å sikre at alle relevante artikler inkluderes (Chai et al., 2021). Tsou et al (2020) mente at menneskelig vurdering av 70-80% trolig var tilstrekkelig for å identifisere alle relevante artikler ved bruk av både Abstrackr og EPPI-Reviewer. De presenterte også et mer risikabelt alternativ hvor kun 50% av artiklene ble vurdert, noe som ifølge deres resultater medførte at 5% av de relevante artiklene ikke ble identifisert. Resultatene fra min oppgave kan på samme måte benyttes til å belyse hvor stor andel av studiene som må menneskelig vurderes før man med sikkerhet kan stole på kalkuleringene fra maskinlæringsverktøyet i Rayyan. Sensitivitet på 100% ble oppnådd etter manuell sortering av 60% av artiklene til denne oppgaven. Basert på dette vil jeg derfor argumentere for at det etter vurdering av 60% potensielt kan være mulig å automatisk ekskludere artiklene med 0.5 og 1.5 stjerner, uten at dette medfører en stor risiko for at relevante artikler ikke identifiseres.

Allerede etter manuell sortering av 20% ble det i min oppgave oppnådd en sensitivitet på 96%. Valizadeh et al (2022) foreslo menneskelig vurdering av 20% som en potensiell grenseverdi å benytte, men alle relevante studier vil da ikke identifiseres. Det er verdt å nevne at ved bruk av tradisjonelle metoder trolig heller ikke oppnås 100% sensitivitet på grunn av menneskelige feil (Thomas et al., 2017). Ved å akseptere noe lavere sensitivitetsverdier vil arbeidsbelastningen ved utarbeiding av systematiske oversiktsartikler reduseres betraktelig, men dette på bekostning av at artiklene da ferdigstilles med et ukjent antall uidentifiserte relevante artikler. Dette er problematisk da systematiske oversiktsartikler i utgangpunktet kjennetegnes av høy grad av metodisk nøyaktighet, og identifikasjon av alle artikler som besvarer den aktuelle problemstillingen. Det er derfor behov for mer forskning for å utforske hvilken betydning artiklene som da ikke identifiseres har for resultatene til den ferdige systematiske oversiktsartikkelen.

6.4.2. Metoder for å forbedre prestasjonen og redusere arbeidsbelastningen

I tillegg til bruk av stoppriterier for å redusere arbeidsbelastningen har flere tidligere studier presentert alternative metoder som potensielt kan forbedre prestasjonen til maskinlæringsverktøyet og dermed redusere arbeidsbelastningen ytterligere.

6.4.2.1. Oversampling

Et alternativ for å forbedre prestasjonen til maskinlæringsvektøyene kan være å skape en høyre andel relevante artikler tidlig i seleksjonsprosessen ved bruk av «oversampling» (Li et al., 2023). Dette gjøres ved at studier man på forhånd vet skal inkluderes sorteres til denne kategorien før resten av seleksjonsprosessen starter. På denne måten får maskinlæringsalgoritmen flere relevante artikler å lære av. Da vil stjerneraneringene trolig bli mer presise, og prestasjonen til maskinlæringsverktøyet vil øke uten en betydelig økning i arbeidsbelastningen. Selv ved bruk av «oversampling» er det likevel verdt å merke seg at punktprevalensen i denne oppgaven ville vært svært lav da den maksimalt når 1.6%. En slik skjevhet i resultatene, med mange falsk positive resultater etter et systematisk søk og dermed få artikler som er aktuelle for inklusjon, kan føre til klassifiseringsproblemer. Dette skyldes at de fleste maskinlæringsalgoritmene er programmert for å maksimere nøyaktigheten, og dermed kan de feilaktig klassifisere relevante studier som irrelevante da dette er det mest sannsynlige utfallet (van Dinter et al., 2021). Det kan tenkes at oversampling potensielt kan redusere risikoen for dette, men effekten er usikker. Det er derfor behov for forskning for å undersøke om dette faktisk forbedrer prestasjonen til maskinlæringsverktøyet.

6.4.2.2. Sortering etter stjerneverdi

Et andre alternativ for å fremskynde seleksjonsprosessen kan være å sortere de resterende artiklene etter stjerneverdi i stede for forfatter etter at de første rangeringene er gjort. En tidligere studie har vist at artiklene med størst betydning for resultatene til den systematiske oversiktsartikkelen fikk gode stjerneverdier (Olofsson et al., 2017). Disse studiene kan derfor identifiseres raskt ved å benytte en høyrere grenseverdi, eller ved å sortere artiklene etter stjerneverdi ved videre studieseleksjon. På denne måten kan relevante studier identifiseres tidligere i studieseleksjonsprosessen, noe som vil gi maskinlæringsalgoritmen mer data å lære fra. Dette kan potensielt også føre til bedre prestasjon ved neste kalkulering. Tidsbruken kan dermed optimaliseres ved at noen forskere kan begynne kritisk vurdering, ekstraksjon av data og analyser når det antas at de fleste og mest relevante studiene er identifisert. Ved gjennomføring av denne oppgaven ble artiklene rangert alfabetisk ved sortering til hvert 20% nivå. Dersom artiklene sorters etter stjerneverdi ved videre vurdering i stedet for i alfabetisk rekkefølge vil dette

trolig føre til at flere relevante artikler identifiseres tidlig i prosessen, og prestasjonen til maskinlæringsverktøyet vil dermed kunne forbedres. Dette eksempelet viser at det på tross av lav sensitivitet med en grenseverdi på ≤2.5 stjerner for eksklusjon, kan være hensiktsmessig å først vurdere de artiklene med høyere stjerneverdi.

6.4.2.3. Redusere menneskelige ressurser

Et siste alternativ som kan trekkes frem for å redusere arbeidsbelastningen ved studieseleksjon er å erstatte den tradisjonelle metoden hvor to uavhengige personer vurderer alle artikler til én som assisteres av et maskinlæringsverktøy (Thomas et al., 2017). Dette er en av de mest lovende metodene for redusert arbeidsbelastning (Olofsson et al., 2017). En studie på effektiviteten av dette ble gjort av Shemilt et al (2016), som konkluderte med at denne metoden kan redusere arbeidsbelastningen med over 60% på bekostning av 5% reduksjon i sensitiviteten. Olofsson et al (2017) trakk også frem at prosessen kan effektiviseres ved å sortere de gjenværende artiklene etter stjernerangeringene fordi relevante artikler identifiseres raskere. De andre forfatterne kan da starte med analyser og videre arbeid tidligere. En kombinasjon av dette og en automatisk eksklusjon av artikler med 0.5 og 1.5 stjerner kan igjen redusere arbeidsbelastningen og kostnadene ytterligere.

For å oppnå en optimal reduksjon i arbeidsbelastningen kan disse metodene også kombineres med bruk av stoppkriterier. Scherhag og Burgarg (2023) trakk i sin artikkel også frem muligheten for en kombinasjon av stoppkriterier for å forbedre prestasjonen ytterligere, for eksempel ved manuell vurdering av 50% for så å fortsette prosessen frem til 50 påfølgende artikler ekskluderes. Ved bruk av Rayyan kan dette for eksempel gjennomføres ved å først inkludere de artiklene man allerede vet skal inkluderes, manuelt vurdere 20% av artiklene etter alfabetisk rekkefølge, sortere etter stjernerangering når de første kalkuleringen er gjennomført for så at en forfatter fortsetter vurderingene til x antall påfølgende studier ekskluderes. For å oppnå tilfredsstillende sensitivitet vil det trolig likevel være behov for å vurdere et relativt høyt antall artikler dersom presisjon til det systematiske søket er lav (Callaghan & Müller-Hansen, 2020). Det er også behov for videre forskning for å fastslå om prestasjonen til Rayyan vil forbedres ved å kombinere stoppkriterier og metoder for å forbedre prestasjonen til maskinlæringsverktøyene, da ingen av de tidligere studiene har undersøkt dette.

6.5. Bruk av maskinlæring for utarbeiding av systematiske oversiktsartikler

På tross av at det allerede er 10 år siden lanseringen av Rayyan og at tidligere forskning har vist lovende resultater vedrørende prestasjonen til maskinlæringsverktøyet brukes programvaren i liten grad ved utarbeiding av systematiske oversiktsartikler i dag (Van Altena et al., 2019). Det hjelper ikke at resultatene fra denne oppgaven og andre studier viser at maskinlæringsalgoritmen i Rayyan og tilsvarende programvarer har en prestasjon som potensielt kan medføre redusert arbeidsbelastning dersom de uavhengig av dette ikke benyttes i praksis. Derfor er det viktig å også utforske potensielle barrierer for bruk av slike programvarer og adressere muligheter for å overvinne disse, da aksept av teknologien er essensielt for suksessfull implementering (Kelly et al., 2023).

6.5.1. Barrierer for implementering og aksept av ny teknologi

Det finnes flere anerkjente teorier for aksept av ny teknologi basert på KI. Tillit, forventet ytelse, forventet innsats, praktisk tilrettelegging, nytteverdi, holdninger og sosial innflytelse trekkes frem som viktige faktorer for aksept av ny teknologi uavhengig av fagfelt (Kelly et al., 2023; Sohn & Kwon, 2020). Tidligere forskning har også utforsket spesifikke barrierer for bruk av programvarer ved utarbeiding av systematiske oversiktsartikler og hvilke faktorer som er essensielle for aksept av slike. O´Connor et al (2019) trakk frem mangel på tillit, utfordringer med oppsett av programvarene, evnen programvarene har til å utføre de aktuelle oppgavene og manglende kunnskap om tilgjengelige programvarer som de største barrierene for implementering. De presiserte også at tilliten til at dagens metoder fører til systematiske oversiktsartikler av høy kvalitet kan lede til en bekymring for at metoder som avviker fra dette fører til studier av lavere kvalitet. Noe av det viktigste for implementering av maskinlæringsverktøy er derfor bevis på at disse presterer like godt eller bedre enn tradisjonelle metoder. Resultatene fra denne oppgaven og tilsvarende studier kan på denne måten bidra til å styrke tiltroen til prestasjonen til programvarene og dermed føre til økt aksept.

Dersom maskinlæringsverktøy skal tas i bruk i praksis er det også essensielt at bruken av de oppleves nyttig og brukervennlig, at de enkelt kan integreres i dagens praksis og at den nye metoden er likeverdig eller bedre enn dagens praksis (O'Connor et al., 2019). Verktøyene må kunne benyttes uten store endringer vedrørende ressurser, programvarer og ferdigheter. Disse faktorene spiller også en

stor rolle ved valg av programvare. Noen av fordelene som trekkes frem ved Rayyan er at programvaren er gratis, ikke krever kunnskap om programmering, er enkel å bruke, kan benyttes til å markere nøkkelord og for å skrive notater knyttet til artiklene (Cleo et al., 2019). På grunn av disse funksjonen kan Rayyan også benyttes uten maskinlæringsverktøyet. Rayyan er derfor et godt hjelpemiddel ved utarbeiding av systematiske oversiktsartikler uavhengig av behovet for et maskinlæringsverktøy. Det at programvaren på denne måten enkelt kan tas i bruk også ved dagens metode for utarbeiding kan bidra til å gjøre implementeringen lettere. Utviklerne av maskinlæringsverktøy for studieseleksjon til systematiske oversiktsartikler bør ha fokus på å utvikle programvarer som er intuitive og enkle å ta i bruk for å oppnå aksept hos brukerne.

Et annet aspekt som kan påvirke menneskers tiltro til maskinlæringsalgoritmene er mangelen på forklaring vedrørende hvilken informasjon konklusjonene til programvaren baserer seg på. Dette er også et av de etiske prinsippene nevnt i metodekapittelet. Ved bruk av maskinlæringsalgoritmen i Rayyan får man for eksempel ikke informasjon om bakgrunnen for stjernevurderingene. Dette påvirker ikke prestasjonen til maskinlæringsalgoritmen, men kan likevel være et viktig tema for diskusjon med tanke på tiltro til programvarene og implementering i praksis. Det er en pågående debatt vedrørende behovet for forklarbarhet ved bruk av KI. Det kan argumenteres for at en godt validert og tilfredsstillende prestasjon er tilstrekkelig, men samtidig finnes det argumenter for integrering av forklarbarhet ved bruk av verktøyene (Amann et al., 2022). Ifølge The European Commission´s High-Level Expert Group on AI er transparens og forklarbarhet viktige faktorer for tillitt til programvarene, og mangel på dette en av de viktigste barrierene for implementering (Amann et al., 2022). Pasienter har rett på en forklaring dersom avgjørelser tas på bakgrunn av informasjon generert av KI (Panch et al., 2018). Ved bruk av veiledet læring hvor utfallet som i dette tilfellet er en binær beslutning, vil det være enkelt for forfatterne å kontrollere avgjørelsene til maskinlæringsverktøyet og behovet for forklaring er derfor ikke like viktig. Mangelen på forklarbarhet setter likevel store krav til validering gjennom undersøkelser av prestasjonen til programvarene (Amann et al., 2022).

Den første artikkelen som ble publisert vedrørende prestasjonen til maskinlæringsverktøyet i Rayyan er allerede åtte år gamle og viste lovende resultater, med en potensiell reduksjon i arbeidsbelastningen på rundt 50% (Ouzzani et al., 2016). Senere studier har også bekreftet at bruk av maskinlæringsverktøyet kan føre til redusert arbeidsbelastning (Olofsson et al., 2017; Valizadeh et al., 2022). På tross av disse resultatene brukes programvaren og tilsvarende verktøy fremdeles i liten grad. En annen potensiell årsak for dette kan være risikoen for at den systematiske oversikten ikke vil bli akseptert av anerkjente tidsskrifter dersom den ikke er basert på tradisjonelle metoder. Som beskrevet i kapittel 6.4.1. risikerer man å ikke identifisere 5% av de relevante artiklene dersom 20% vurderes manuelt. Det er ingen konsensus rundt hvor stor potensiell reduksjon i sensitivitet som aksepteres. Fordelene bruk av maskinlæringsverktøyet tilfører, med redusert arbeidsbelastning, kostnad og tidsbruk, vil da ikke veie opp for risikoen for at studien ikke publiseres. For vellykket implementering av maskinlæringsverktøy er det derfor, i tillegg til mer forskning på prestasjonen til programvarene, viktig at anerkjente tidsskrifter også aksepterer dette som en metode som leder til resultater med høy validitet og reliabilitet. Samt at det utarbeides retningslinjer for hvordan disse skal benyttes. Det er da behov for videre forskning for å fastslå at programvarene presterer tilfredsstillende og metoder for hvordan de skal brukes.

6.5.3 Fremtiden for maskinlæringsverktøy til utarbeiding av systematiske oversiktsartikler

Bruken av maskinlæringsverktøy har et stort potensial, og implementering av KI ved utarbeiding av systematiske oversiktsartikler kan forkorte tiden det tar før forskningsbasert kunnskap implementeres i klinisk praksis. Som vist av resultatene kan arbeidsbelastningen, og dermed tidsbruken, ved utarbeiding av systematiske oversiktsartikler reduseres ved å benytte maskinlæringsalgoritmer ved seleksjon av studier. Dette er et viktig steg i riktig retning, men det optimale målet er likevel helautomatisering av prosessen ved hjelp av programvarer som kan gjennomføre selvstendig utarbeiding og kontinuerlig oppdatering av systematiske oversiktsartikler. Ouzzani (2016) trekker også frem dette som det endelige målet ved utviklingen av Rayyan. Resultatene fra både tidligere og denne studien viser stort potensiale for redusert arbeidsbelastning for studieseleksjonsprosessen, men automatisering på

nivå 3 og 4 (O'Connor et al., 2019) ser enda ikke ut til å være realistisk med denne programvaren og mer avanserte former for KI er trolig nødvendig for å oppnå dette.

Programvaren Research Screener bruker som sagt mer avanserte former for KI ved studieseleksjon til systematiske oversiktsartikler, og har vist lovende resultater (Chai et al., 2021). Også K.R. Felizardo har publisert flere artikler (2012; 2014; 2011) som undersøker bruken av uveiledet læring istedenfor veiledet læring til utarbeiding av systematiske oversiktsartikler. I disse artiklene ble «Visual text mining» vurdert som et hensiktsmessig verktøy for å selektere studier og oppdatere systematiske oversiktsartikler. Denne teknologien ble også brukt til studieseleksjon av Li et al (2016) som fant en redusert arbeidsbelastning på 91.8, 85.7 og 49.3% ved en sensitivitet på 100%. Natukunda et al (2023) presenterte også en metode for studieseleksjon basert på uveiledet læring. Selv om sensitiviteten i denne studien var lav konkluderte de med at uveiledet læring har potensiale til å kunne redusere arbeidsbelastningen og at automatisering av prosessen virker overkommelig ved bruk av denne teknologien. Bruken av dyplæring for å støtte utarbeidingen av kontinuerlig oppdaterte systematiske oversiktsartikler ble også undersøkt i 2023 i forbindelse med Covid-19 pandemien, som virkelig tydeliggjorde betydningen av hurtig utarbeiding av forskning i det medisinske fagfeltet. Fremgangsmetoden som ble brukt i denne studien oppnådde en sensitivitet på 89% og forfatterne konkluderte med at studien viste at dyplæring har potensiale til å semi-automatisere prosessen (Knafou et al., 2023). Både veiledet og uveiledet læring, inkludert dyplæring, viser dermed potensiale for å redusere arbeidsbelastningen ved utarbeiding av systematiske oversiktsartikler. Prestasjonen til maskinlæringsverktøyene basert på uveiledet læring ser derimot ikke ut til å være tydelig overlegne de resultatene som ble oppnådd ved bruk av veiledet læring. Uavhengig av valg av teknologi ser det derfor fremdeles ut til å være behov for menneskelige vurderinger og dermed enda ikke aktuelt med fullstendig automatisering på nivå 4.

En annen av utfordringene ved å implementere bruk av maskinlæringsverktøy ved utarbeiding av systematiske oversiktsartikler er nettopp det at mange av verktøyene er utviklet uavhengig av hverandre og derfor ikke er kompatible (Tsafnat et al., 2014). Det er i dag derfor behov for å benytte flere ulike verktøy for de ulike stadiene av utarbeidingen. Rayyan kan slik programvaren er i dag ikke benyttes til andre steg

enn studieseleksjon. Dersom flere av stadiene skal automatiseres vil det derfor i dag være nødvendig å benytte flere ulike programvarer til ulike formål. Om målet om helautomatisering av prosessen og kontinuerlig oppdatering av systematiske oversiktsartikler skal være realistisk er det helt nødvendig med fokus på å utvikle verktøy som kan gjennomføre alle stadier av prosessen. Dersom samtlige av disse stadiene automatiseres åpner det også opp muligheten for kontinuerlig automatisk oppdatering, hvor artiklene oppdateres automatisk så fort ny relevant kunnskap blir tilgjengelig (Akl et al., 2017). Ved å kontinuerlig oppdatere systematiske oversiktsartikler, kan funn av effektive behandlingsmetoder raskere implementeres i klinisk praksis (Mulrow, 1994). Det å kontinuerlig oppdatere systematiske oversiktsartikler vil også effektivisere prosessen med å utvikle kliniske retningslinjer (Akl et al., 2017). Dersom denne prosessen i tillegg kan automatiseres vil implementeringsprosessen forkortes ytterligere, og fysioterapeuter i klinisk praksis vil alltid ha tilgang på nyeste forskning. Dette kan potensielt gjøre det lettere å holde seg faglig oppdater og dermed bedre pasientbehandlingen.

6.6. Metodediskusjon

For å gjennomføre denne metodestudien ble maskinlæringsverktøyets evne til å identifisere artikler som er relevante for inklusjon sammenlignet med menneskelige vurderinger. Dette ble gjort for å undersøke om programvaren brukes for å optimalisere og effektivisere studieseleksjonsprosessen ved utarbeidelsen av en systematisk oversiktsartikkel. På bakgrunn av tidligere forskning valgte jeg å vurdere prestasjonen ved to forskjellige grenseverdier på <2.5 og ≤2.5 stjerner for eksklusjon. Basert på resultatene viste dette seg å være hensiktsmessige grenseverdier å benytte, da høyere eller lavere grenseverdier ikke ville ført til bedre prestasjon. Styrker og begrensninger ved metoden som ble brukt i denne studien vil bli presentert i de kommende avsnittene, og kan dermed også danne grunnlag for anbefalinger for videre studier.

6.6.1. Referansestandard

Det er opplagt innslag av subjektivitet og skjønn ved seleksjon av studier til systematiske oversiktsartikler ved bruk av tradisjonelle metoder basert på menneskelige vurderinger (FHI, 2022). Det er disse vurderingene benyttes som gullstandard både i min oppgave og tilsvarende studier. Det er derfor aktuelt å spørre seg i hvilken grad vi kan være sikre på at denne gullstandarden er riktig, og

eventuelle fordeler og ulemper ved at subjektiviteten i avgjørelsene elimineres. Et poeng som er verdt å diskutere når man sammenlignet bruken av maskinlæringsverktøy med menneskelige vurderinger er at man ikke nødvendigvis oppnår perfekt sensitivitet ved tradisjonelle metoder heller (Thomas et al., 2017). Selv om seleksjonen av studier til den systematiske oversiktsartikkelen som ble brukt til denne oppgaven ble utført av minst to erfarne forskere kan det ikke utelukkes at studier ble feilaktig inkludert og ekskludert. Ved bruk av veiledet læring, som ved bruk av maskinlæringsalgoritmen i Rayyan, vil disse subjektive avgjørelsene påvirke avgjørelsene og dermed prestasjonen til maskinlæringsalgoritmen. Ved vurdering av prestasjonen til maskinlæringsverktøyet i denne oppgaven er det også en forutsetning at artiklene sorteres til riktig kategori. Menneskelig oppmerksomhet svikter ved repetitivt arbeid, noe som gjør studieseleksjonsprosessen sårbar for feil (van Dinter et al., 2021). Hvis man på grunn av dårlig konsentrasjon sorterer treningsdataene til feil kategori, vil det medføre at maskinen vil lære det motsatte av det den skal. Treningsdatasettet ville dermed fått skjevheter som følge av menneskelige feilvurderinger. Imidlertid ble det ved hvert 20%-stadium verifisert at det antall inkluderte studier samsvaret med referansestandarden, som skal ha eliminert dette som en feilkilde.

6.6.3. Valg av systematisk oversiktsartikkel

Ingen av de tidligere identifiserte studiene omhandler muskelskjeletthelse eller prognostiske studier. Det er derfor en styrke at min oppgave gir ny innsikt vedrørende prestasjonen til Rayyan på dette området. Tidligere forskning har vist at klassifiserere basert på maskinlæring presterer bedre ved seleksjon av randomiserte kontrollerte studier (Thomas et al., 2017; Wallace et al., 2017). Denne oppgaven gir derfor viktig innsikt i prestasjonen til Rayyan ved studieseleksjon til en prognostisk systematisk oversiktsartikkel hvor også prospektive kohortstudier og registerbaserte studier var aktuelle for inklusjon. I en tidligere studie som vurderte sensitiviteten og spesifisiteten til maskinlæringsverktøyet i Rayyan ble det trukket frem som en begrensing at de kun hadde 500 treff etter det systematiske søket, og at mer presise estimater for sensitivitet og spesifisitet kunne blitt oppnådd ved bruk av et større datasett (Li et al., 2023). Til sammenligning ble det i denne studien benyttet et datasett på 7994 artikler. Det kan på bakgrunn av dette antas at min studie førte til mer presise estimater og dermed i større grad reliable resultater.

Det må likevel nevnes som en begrensning ved denne oppgaven at jeg kun undersøkte prestasjonen til maskinlæringsverktøyet for studieseleksjon til en enkelt systematisk oversiktsartikkel. Selv om det ble benyttet et stort datasett er det derfor vanskelig å vurdere om resultatene er generaliserbare til andre oversiktsartikler, som for eksempel omhandler andre tema, har et ulikt antall treff eller annen presisjon etter det systematiske litteratursøket. Flere av de tidligere studiene har benyttet flere oversiktsartikler og sett på gjennomsnittet av resultatene, dette kan potensielt styrke verdien av resultatene og gjøre det lettere å avgjøre om de er generaliserbare. Det faktum at resultatene likevel samsvarer i stor grad med tidligere forskning kan tyde på at prestasjonen er relativt stabil uavhengig av tema. Ved at ulike studier tester prestasjonen på et bredt spekter av ulike tema kan dette samlet sett føre til økt generaliserbarhet og tiltro til maskinlæringsalgoritmen. Samtidig kan det også gi en indikasjon på om prestasjonen er akseptabel eller suboptimal for ulike studiedesign og studiepopulasjoner, og dermed om det vil være aktuelt å benytte maskinlæring for studieseleksjon.

6.6.2. Valg av utfallsmål

Resultatene i denne oppgaven ble presentert ved å kalkulere sensitivitet, spesifisitet, PPV og NPV. En tidligere studie trekker frem nettopp disse utfallsmålene som relevante for vurdering av prestasjonen til maskinlæringsverktøy ved utarbeiding av systematiske oversiktsartikler (O'Connor et al., 2019). Dersom ulike studier benytter ulike utfallsmål og vil det være utfordrende å sammenligne resultatene. Dersom det i fremtidige studier benyttes samme utfallsmål for å presentere resultatene kan dette etter hvert samles i en systematisk oversiktsartikkel og metaanalyse. Dette vil gjøre det lettere å trekke beslutninger vedrørende generaliserbarheten og prestasjonen til Rayyan. Tidligere studier har, i tillegg til å vurdere evnen maskinlæringsverktøyet har til å vurdere titler og sammendrag, sett på prestasjonen til Rayyan vedrørende vurdering av artiklene i fulltekst. Dette ble ikke gjennomført i denne oppgaven.

En begrensning ved utfallmålene som ble rapportert i denne oppgaven er at den potensielle reduksjonen i tid, og dermed kostnader, brukt på seleksjonsprosessen ikke ble vurdert. Dette skyldes at studieseleksjonen til den systematiske oversiktsartikkelen som benyttes som referansestandard allerede var gjennomført ved oppstart av denne oppgaven. Det var derfor ikke mulig å skaffe data som

eventuelt kunne blitt brukt som sammenligningsgrunnlag. Dette er en tydelig svakhet da redusert tidsbruk som nevnt er hovedformålet med automatisering av studieseleksjonen (van Dinter et al., 2021). Resultatene fra en studie publisert i 2016 viste at studieseleksjon ved bruk av et maskinlæringsverktøy var opptil 98% mer kostnadseffektivt sammenlignet med tradisjonelle metoder. Det er viktig å merke seg at bruk av denne metoden samtidig gikk negativt ut over sensitiviteten, som ble redusert med 5% (Shemilt et al.). Fremtidige studier kan med fordel inkludere redusert tidsbruk og kostnadseffektivitet som utfallsmål, i tillegg til de nevnte utfallsmålene sensitivitet, spesifisitet, NPV og PPV. Da både prestasjon og redusert bruk av tid er viktige faktorer for verdien verktøyene kan tilføre.

Det er også en begrensning at jeg i denne oppgaven ikke vurderte betydningen de inkluderte artiklene fikk for resultatene til den systematiske oversiktsartikkelen i sammenheng med stjernerangeringene fra Rayyan. Heller ikke dette var mulig å gjennomføre i mitt prosjekt, da den systematiske oversiktsartikkelen fremdeles er under utarbeiding og resultatene dermed ikke er publisert. Gjennomføring av dette ville vært nyttig for å se om resultatene samsvarte med de fra Olofsson et al (2017), som viste at de studiene med lav eller moderat risiko for feil ble identifisert tidlig i seleksjonsprosessen. Dersom disse studiene identifiseres tidlig kan det argumenteres for at en liten reduksjon i sensitivitet ikke vil være like avgjørende for det endelige resultatet i en systematisk oversiktsartikkel og at lavere sensitivitetsverdier ved studieseleksjon dermed kan aksepteres.

6.6.4 Et fagfelt i hurtig endring

En siste begrensning jeg vil trekke frem er at dette er et fagfelt i hurtig endring. Det kommer stadig nye programvarer og det gjøres endringer i de eksisterende programvarene. Rayyan har gjennomgått større endringer etter gjennomføring av studieseleksjonen til denne oppgaven. Dette kan trolig ha stor påvirkning på opplevelsen og brukervennligheten til programvaren, noe som har blitt trukket frem som fordeler ved Rayyan i tidligere studier. I tillegg til stjernerangeringene som ble benyttet i denne studien har det nå blitt lagt til en anbefaling relatert til stjernerangeringen. Disse er som følger: 4.5 stjerner = mest sannsynlig relevant for inklusjon, 3.5 stjerner = sannsynligvis relevant for inklusjon, 2.5 stjerner = ingen anbefaling, 1.5 stjerner = sannsynligvis ikke relevant for inklusjon, 0.5 stjerne = mest

sannsynlig ikke relevant for inklusjon. Denne anbefalingen styrker behovet for videre forskning på prestasjonen til maskinlæringsverktøyet rundt en stjerneverdi på 2.5 stjerner for eksklusjon, da det per nå ikke klarer å gi en anbefaling for disse artiklene. Rayyan har også fått en AI-assistent og flere funksjoner som kan benyttes mot betaling. Dersom det gjøres endringer i selve maskinlæringsalgoritmen vil dette kunne påvirke prestasjonen til Rayyan ved studieseleksjon. Imidlertid er det ikke gitt ut informasjon om endringer av denne og resultatene vedrørende prestasjonen til maskinlæringsverktøyet vil derfor fremdeles være gjeldene.

6.6.5. Etiske betraktninger:

Vi møter på mange utfordringer når maskinlæringsverktøy skal implementeres i det medisinske fagfeltet, men særlig to faktorer fører til store begrensninger. Både ved bruk i klinisk praksis og i forskning håndteres det store mengder sensitive personopplysninger. Det er strenge krav til innsamling, lagring og bruk av slik informasjon, noe som begrenser hvilken informasjon programvarene kan få tilgang til og dermed hvor gode modellene kan bli (Strümke, 2023). Som beskrevet i metode kapittelet gir bruk av KI til forskning i det medisinske fagfeltet i mange tilfeller programvarene tilgang på store mengder personopplysninger og sensitive data. Dette er sjelden en problemstilling ved utarbeiding av systematiske oversiktsartikler, da dataene primært baserer seg på gjennomsnittverdier på gruppenivå og spredningsmål, uten direkte personidentifiserbare opplysninger. Bruk av anonymiserte data trekkes i den nasjonale strategien frem som et mer personvennlig alternativ. Jeg vil derfor argumentere for at utarbeiding av systematiske oversiktsartikler er et gunstig sted å begynne implementeringen av løsninger baser på KI i helsevesenet.

7. Konklusjon

Lav sensitivitet ved alle nivåer av studieseleksjonen ved en grenseverdi på ≤ 2.5 stjerner for eksklusjon tyder på at en betydelig andel relevante studier ville blitt feilaktig ekskludert ved denne grenseverdien. Ved en grenseverdi på <2.5 for eksklusjon har maskinlæringsverktøyet god evne til å ekskludere irrelevante artikler, men relativt dårlig evne til å identifisere artiklene som skal inkluderes. Basert på mine resultater vil jeg argumentere for at det etter vurdering av 60% potensielt kan være mulig å ekskludere artiklene med 0.5 og 1.5 stjerner uten ytterligere vurdering, uten at dette medfører en stor risiko for at relevante artikler ikke identifiseres. Allerede etter manuell vurdering av 20% av artiklene ble det oppnådd en sensitivitetsverdi på over 95%. Menneskelig vurdering av kun 20% av artiklene kan dermed være et alternativ i situasjoner hvor det ikke er essensielt å identifisere alle relevante artikler. Det integrerte maskinlæringsverktøyet i Rayyan presterer dermed godt nok til å kunne redusere arbeidsbelastningen ved studieseleksjon til en systematisk oversiktsartikkel på prognostiske modeller for degenerativ ryggkirurgi, men menneskelige vurderinger er fremdeles i stor grad nødvendig og videre utvikling av verktøyet er essensielt før full automatisering av oppgaven.

8. Referanseliste

AI, H. (2019). High-level expert group on artificial intelligence. *Ethics guidelines for trustworthy AI*, *6*.

Akl, E. A., Meerpohl, J. J., Elliott, J., Kahale, L. A., Schünemann, H. J., Agoritsas, T., Hilton, J., Perron, C., Akl, E., & Hodder, R. (2017). Living systematic reviews: 4. Living guideline recommendations. *Journal of clinical epidemiology*, *91*, 47-53.

Amann, J., Vetter, D., Blomberg, S. N., Christensen, H. C., Coffee, M., Gerke, S., Gilbert, T. K., Hagendorff, T., Holm, S., & Livne, M. (2022). To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health*, *1*(2), e0000016.

Balas, E. A., & Boren, S. A. (2000). Managing clinical knowledge for health care improvement. *Yearbook of medical informatics*, *9*(01), 65-70.

Bastian, H., Glasziou, P., & Chalmers, I. (2010). Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine*, *7*(9), e1000326.

Blaizot, A., Veettil, S. K., Saidoung, P., Moreno-Garcia, C. F., Wiratunga, N., Aceves-Martins, M., Lai, N. M., & Chaiyakunapruk, N. (2022). Using artificial intelligence methods for systematic review in health sciences: A systematic review. *Research synthesis methods*, *13*(3), 353-362.

Borah, R., Brown, A. W., Capers, P. L., & Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ open*, *7*(2), e012545.

Callaghan, M. W., & Müller-Hansen, F. (2020). Statistical stopping criteria for automated screening in systematic reviews. *Systematic reviews*, *9*, 1-14.

Chai, K. E., Lines, R. L., Gucciardi, D. F., & Ng, L. (2021). Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews. *Systematic reviews*, *10*, 1-13.

Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *The Lancet*, *374*(9683), 86-89.

Chung, G., & Coiera, E. (2007). A study of structured clinical abstracts and the semantic classification of sentences. Biological, translational, and clinical language processing,

Cierco Jimenez, R., Lee, T., Rosillo, N., Cordova, R., Cree, I. A., Gonzalez, A., & Indave Ruiz, B. I. (2022). Machine learning computational tools to assist the performance of systematic reviews: A mapping review. *BMC Medical Research Methodology*, *22*(1), 1-14.

Clark, J., Glasziou, P., Del Mar, C., Bannach-Brown, A., Stehlik, P., & Scott, A. M. (2020). A full systematic review was completed in 2 weeks using automation tools: a case study. *Journal of clinical epidemiology*, *121*, 81-90.

Cleo, G., Scott, A. M., Islam, F., Julien, B., & Beller, E. (2019). Usability and acceptability of four systematic review automation software packages: a mixed method design. *Systematic reviews*, *8*(1), 1-5.

Cohen, A. M., Ambert, K., & McDonagh, M. (2012). Studying the potential impact of automated document classification on scheduling a systematic review update. *BMC medical informatics and decision making*, *12*, 1-11.

Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P.-Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, *13*(2), 206-219.

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future healthcare journal*, *6*(2), 94.
Di Nunzio, G. M. (2018). A study of an automatic stopping strategy for technologically assisted medical reviews. Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40,
Dickersin, K., Scherer, R., & Lefebvre, C. (1994). Systematic reviews: identifying relevant studies for systematic reviews. *Bmj*, *309*(6964), 1286-1291.
Elliott, J. H., Synnot, A., Turner, T., Simmonds, M., Akl, E. A., McDonald, S., Salanti, G., Meerpohl, J., MacLehose, H., & Hilton, J. (2017). Living systematic review: 1. Introduction—the why, what, when, and how. *Journal of clinical epidemiology*, *91*, 23-30.
Elliott, J. H., Turner, T., Clavisi, O., Thomas, J., Higgins, J. P., Mavergames, C., & Gruen, R. L. (2014). Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS medicine*, *11*(2), e1001603.
Felizardo, K. R., Andery, G. F., Paulovich, F. V., Minghim, R., & Maldonado, J. C. (2012). A visual analysis approach to validate the selection review of primary studies in systematic reviews. *Information and Software Technology*, *54*(10), 1079-1091.
Felizardo, K. R., Nakagawa, E. Y., MacDonell, S. G., & Maldonado, J. C. (2014). A visual analysis approach to update systematic reviews. Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering,
Felizardo, K. R., Riaz, M., Sulayman, M., Mendes, E., MacDonell, S. G., & Maldonado, J. C. (2011). Analysing the use of graphs to represent the results of systematic reviews in software engineering. 2011 25th Brazilian Symposium on Software Engineering.
FHI. (2022, 22. April). *Slik oppsummerer vi forskning*. FHI. https://www.fhi.no/ku/oppsummert-forskning-for-helsetjenesten/metodeboka/framgangsmate/hente-ut-data-sammenfatte-og-gradere/?term=
Gaskins, N. J., Bray, E., Hill, J. E., Doherty, P. J., Harrison, A., & Connell, L. A. (2021). Factors influencing implementation of aerobic exercise after stroke: a systematic review. *Disability and rehabilitation*, *43*(17), 2382-2396.
Gates, A., Guitard, S., Pillay, J., Elliott, S. A., Dyson, M. P., Newton, A. S., & Hartling, L. (2019). Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools. *Systematic reviews*, *8*(1), 1-11.
Gates, A., Johnson, C., & Hartling, L. (2018). Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the Abstrackr machine learning tool. *Systematic reviews*, *7*, 1-9.
Goldet, G., & Howick, J. (2013). Understanding GRADE: an introduction. Journal of Evidence-Based Medicine, 6(1), 50-54.
Greenhalgh, T. (2019). *How to read a paper : the basics of evidence-based medicine and healthcare* (Sixth edition. ed.). John Wiley & Sons Ltd.
Haddaway, N. R., & Westgate, M. J. (2019). Predicting the time needed for environmental systematic reviews and systematic maps. *Conservation Biology*, *33*(2), 434-443.

Harrison, H., Griffin, S. J., Kuhn, I., & Usher-Smith, J. A. (2020). Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. *BMC Medical Research Methodology*, *20*, 1-12.
Helsebiblioteket. (2017, 28. august). *Kildevalg.* Helsebiblioteket. https://www.helsebiblioteket.no/innhold/artikler/kunnskapsbasert-praksis/kunnskapsbasertpraksis.no.
Helsedirektoratet. (2022, 16. mars). *Rammer og retning for kunstig intelligens.* Helsedirektoratet. https://www.helsedirektoratet.no/tema/kunstig-intelligens
Helsepersonelloven. (1999). *Lov om helsepersonell (LOV-1999-07-02-64).* Lovdata. https://lovdata.no/lov/1999-07-02-64/§4
Higgins, J. P., & Green, S. (2008). Cochrane handbook for systematic reviews of interventions.
Hsu, W., Speier, W., & Taira, R. K. (2012). Automated extraction of reported statistical analyses: towards a logical representation of clinical trial literature. AMIA Annual Symposium Proceedings,
Jamtvedt, G., Hagen, K., & Bjørndal, A. (2003). Kunnskapsbasert fysioterapi. *Metoder og arbeidsmåter*, *1*.
Jaspers, S., De Troyer, E., & Aerts, M. (2018). Machine learning techniques for the automation of literature reviews and systematic reviews in EFSA. *EFSA Supporting Publications*, *15*(6), 1427E.
Jonnalagadda, S., & Petitti, D. (2013). A new iterative method to reduce workload in systematic review process. *International journal of computational biology and drug design*, *6*(1-2), 5-17.
Jonnalagadda, S. R., Goyal, P., & Huffman, M. D. (2015). Automating data extraction in systematic reviews: a systematic review. *Systematic reviews*, *4*(1), 1-16.
Kelly, S., Kaye, S.-A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*, *77*, 101925.
Kiritchenko, S., De Bruijn, B., Carini, S., Martin, J., & Sim, I. (2010). ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC medical informatics and decision making*, *10*, 1-17.
Knafou, J., Haas, Q., Borissov, N., Counotte, M., Low, N., Imeri, H., Ipekci, A. M., Buitrago-Garcia, D., Heron, L., & Amini, P. (2023). Ensemble of deep learning language models to support the creation of living systematic reviews for the COVID-19 literature. *Systematic reviews*, *12*(1), 94.
Koch, G. (2006). No improvement–still less than half of the Cochrane reviews are up to date. XIV Cochrane Colloquium, Dublin, Ireland.
Kommunal- og moderniseringsdepartementet. (2020, 14. januar). *Nasjonal strategi for kunstig intelligens.* Regjeringen. https://www.regjeringen.no/no/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/
Lefebvre, C., Glanville, J., Briscoe, S., Littlewood, A., Marshall, C., Metzendorf, M. I., Noel-Storr, A., Rader, T., Shokraneh, F., & Thomas, J. (2019). Searching for and selecting studies. *Cochrane Handbook for systematic reviews of interventions*, 67-107.
Lefebvre, C., Glanville, J., Wieland, L. S., Coles, B., & Weightman, A. L. (2013). Methodological developments in searching for studies for systematic reviews: past, present and future? *Systematic reviews*, *2*(1), 1-9.

Lensen, S., Farquhar, C., & Jordan-Cole, V. (2014). Risk of bias: are judgements consistent between reviews? 22nd Cochrane Colloquium,
Li, D., Wang, Z., Wang, L., Sohn, S., Shen, F., Murad, M. H., & Liu, H. (2016). A text-mining framework for supporting systematic reviews. *American journal of information management*, *1*(1), 1.
Li, J., Kabouji, J., Bouhadoun, S., Tanveer, S., Filion, K. B., Gore, G., Josephson, C. B., Kwon, C.-S., Jette, N., & Bauer, P. R. (2023). Sensitivity and specificity of alternative screening methods for systematic reviews using text mining tools. *Journal of clinical epidemiology*, *162*, 72-80.
Marshall, C. (2016). *Tool support for systematic reviews in software engineering* Keele University].
Marshall, I. J., Kuiper, J., & Wallace, B. C. (2014). Automating risk of bias assessment for clinical trials. proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics,
Marshall, I. J., & Wallace, B. C. (2019). Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic reviews*, *8*, 1-10.
Millard, L. A., Flach, P. A., & Higgins, J. P. (2016). Machine learning to assist risk-of-bias assessments in systematic reviews. *International journal of epidemiology*, *45*(1), 266-277.
Miwa, M., Thomas, J., O'Mara-Eves, A., & Ananiadou, S. (2014). Reducing systematic review workload through certainty-based screening. *Journal of biomedical informatics*, *51*, 242-253.
Mueller, J. P., & Massaron, L. (2021). *Machine learning for dummies*. John Wiley & Sons.
Mulrow, C. D. (1994). Systematic reviews: rationale for systematic reviews. *Bmj*, *309*(6954), 597-599.
Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, *18*, 1-7.
Nascimento, J. d. S. G., Siqueira, T. V., Oliveira, J. L. G. d., Alves, M. G., Regino, D. d. S. G., & Dalri, M. C. B. (2021). Development of clinical competence in nursing in simulation: the perspective of Bloom's taxonomy. *Revista Brasileira de Enfermagem*, *74*.
Natukunda, A., & Muchene, L. K. (2023). Unsupervised title and abstract screening for systematic review: a retrospective case-study using topic modelling methodology. *Systematic reviews*, *12*(1), 1.
Navarro, C. L. A., Damen, J. A., Takada, T., Nijman, S. W., Dhiman, P., Ma, J., Collins, G. S., Bajpai, R., Riley, R. D., & Moons, K. G. (2021). Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *Bmj*, *375*.
Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, *24*(12), 1565-1567.
O'Connor, A. M., Tsafnat, G., Thomas, J., Glasziou, P., Gilbert, S. B., & Hutton, B. (2019). A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? *Systematic reviews*, *8*(1), 1-8.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, *4*(1), 1-22.
Olofsson, H., Brolund, A., Hellberg, C., Silverstein, R., Stenström, K., Österberg, M., & Dagerhamn, J. (2017). Can abstract screening workload be reduced using text mining? User experiences of the tool Rayyan. *Research synthesis methods*, *8*(3), 275-280.
OpenAI. (2022, 30. november). *Introducing ChatGPT.* OpenAI. https://openai.com/index/chatgpt/
Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, *5*, 1-10.
Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Bmj, 372.
Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health systems. *Journal of global health*, *8*(2).
Pinna, F., Manchia, M., Paribello, P., & Carpiniello, B. (2020). The impact of alexithymia on treatment response in psychiatric disorders: a systematic review. *Frontiers in Psychiatry*, *11*, 311.
Przybyła, P., Brockmeier, A. J., Kontonatsios, G., Le Pogam, M. A., McNaught, J., von Elm, E., Nolan, K., & Ananiadou, S. (2018). Prioritising references for systematic reviews with RobotAnalyst: a user study. *Research synthesis methods*, *9*(3), 470-488.
PubMed. (2024). *PubMed.* NIH. https://pubmed.ncbi.nlm.nih.gov/
Qureshi, R., Shaughnessy, D., Gill, K. A., Robinson, K. A., Li, T., & Agai, E. (2023). Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Systematic reviews*, *12*(1), 72.
Regjeringen. (2020). *Nasjonal strategi for kunstig intelligens* Departementenes sikkerhets- og serviceorganisasjon Retrieved from https://www.regjeringen.no/contentassets/1febbbb2c4fd4b7d92c67ddd353b6ae8/no/pdfs/ki-strategi.pdf
Rogers, C. R., Matthews, P., Xu, L., Boucher, K., Riley, C., Huntington, M., Le Duc, N., Okuyemi, K. S., & Foster, M. J. (2020). Interventions for increasing colorectal cancer screening uptake among African-American men: a systematic review and meta-analysis. *PLoS One*, *15*(9), e0238354.
Scherhag, J., & Burgard, T. (2023). Performance of semi-automated screening using Rayyan and ASReview: A retrospective analysis of potential work reduction and different stopping rules. *Big Data & Research Syntheses 2023, Frankfurt, Germany*.
Schünemann, H. J., & Moja, L. (2015). Reviews: rapid! rapid! rapid!… and systematic. *Systematic reviews*, *4*(1), 1-3.
Shemilt, I., Khan, N., Park, S., & Thomas, J. (2016). Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Systematic reviews*, *5*, 1-13.
Shemilt, I., Simon, A., Hollands, G. J., Marteau, T. M., Ogilvie, D., O'Mara-Eves, A., Kelly, M. P., & Thomas, J. (2014). Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research synthesis methods*, *5*(1), 31-49.

Shojania, K. G., Sampson, M., Ansari, M. T., Ji, J., Doucette, S., & Moher, D. (2007). How quickly do systematic reviews go out of date? A survival analysis. *Annals of internal medicine*, *147*(4), 224-233.
Sohn, K., & Kwon, O. (2020). Technology acceptance theories and factors influencing artificial Intelligence-based intelligent products. *Telematics and Informatics*, *47*, 101324.
Strümke, I. (2023). *Maskiner som tenker: algoritmenes hemmeligheter og veien til kunstig intelligens*. Kagge forlag.
Summerscales, R., Argamon, S., Hupert, J., & Schwartz, A. (2009). Identifying treatments, groups, and outcomes in medical abstracts. The Sixth Midwest Computational Linguistics Colloquium (MCLC 2009),
Thomas, J. (2013). Diffusion of innovation in systematic review methodology: why is study selection not yet assisted by automation. *OA Evidence-Based Medicine*, *1*(2), 1-6.
Thomas, J., McNaught, J., & Ananiadou, S. (2011). Applications of text mining within systematic reviews. *Research synthesis methods*, *2*(1), 1-14.
Thomas, J., Noel-Storr, A., Marshall, I., Wallace, B., McDonald, S., Mavergames, C., Glasziou, P., Shemilt, I., Synnot, A., & Turner, T. (2017). Living systematic reviews: 2. Combining human and machine effort. *Journal of clinical epidemiology*, *91*, 31-37.
Torres Torres, M., & Adams, C. E. (2017). RevManHAL: towards automatic text generation in systematic reviews. *Systematic reviews*, *6*, 1-7.
Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., & Coiera, E. (2014). Systematic review automation technologies. *Systematic reviews*, *3*, 1-15.
Tsou, A. Y., Treadwell, J. R., Erinoff, E., & Schoelles, K. (2020). Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer. *Systematic reviews*, *9*, 1-14.
Valizadeh, A., Moassefi, M., Nakhostin-Ansari, A., Hosseini Asl, S. H., Saghab Torbati, M., Aghajani, R., Maleki Ghorbani, Z., & Faghani, S. (2022). Abstract screening using the automated tool rayyan: Results of effectiveness in three diagnostic test accuracy systematic reviews. *BMC Medical Research Methodology*, *22*(1), 1-15.
Van Altena, A., Spijker, R., & Olabarriaga, S. (2019). Usage of automation tools in systematic reviews. *Research synthesis methods*, *10*(1), 72-82.
Van De Schoot, R., De Bruin, J., Schram, R., Zahedi, P., De Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., & Ferdinands, G. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature machine intelligence*, *3*(2), 125-133.
van Dinter, R., Tekinerdogan, B., & Catal, C. (2021). Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology*, *136*, 106589.
Wallace, B. C., Noel-Storr, A., Marshall, I. J., Cohen, A. M., Smalheiser, N. R., & Thomas, J. (2017). Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. *Journal of the American Medical Informatics Association*, *24*(6), 1165-1168.
Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C., & Schmid, C. H. (2010). Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics*, *11*(1), 1-11.

Yu, F., Liu, C., & Sharmin, S. (2022). Performance, usability, and user experience of rayyan for systematic reviews. *Proceedings of the Association for Information Science and Technology*, *59*(1), 843-844.
Yu, Z., & Menzies, T. (2019). FAST2: An intelligent assistant for finding relevant papers. *Expert Systems with Applications*, *120*, 57-71.

9. Vedlegg

Krysstabeller fra SPSS:

star_20 * decision_all Crosstabulation

			decision_all		
			Excluded	Included	Total
star_20	.5	Count	1714	2	1716
		% within decision_all	27.3%	1.9%	26.9%
	1.5	Count	675	2	677
		% within decision_all	10.7%	1.9%	10.6%
	2.5	Count	3890	101	3991
		% within decision_all	61.9%	96.2%	62.5%
	3.5	Count	1	0	1
		% within decision_all	0.0%	0.0%	0.0%
Total		Count	6280	105	6385
		% within decision_all	100.0%	100.0%	100.0%

star_40 * decision_all Crosstabulation

			decision_all		
			Excluded	Included	Total
star_40	.5	Count	1554	0	1554
		% within decision_all	32.9%	0.0%	32.4%
	1.5	Count	1077	1	1078
		% within decision_all	22.8%	1.3%	22.5%
	2.5	Count	2084	62	2146
		% within decision_all	44.2%	80.5%	44.7%
	3.5	Count	4	14	18
		% within decision_all	0.1%	18.2%	0.4%
Total		Count	4719	77	4796
		% within decision_all	100.0%	100.0%	100.0%

star_60 * decision_all Crosstabulation

			decision_all		
			Excluded	Included	Total
star_60	.5	Count	1313	0	1313
		% within decision_all	42.0%	0.0%	41.4%
	1.5	Count	749	0	749
		% within decision_all	23.9%	0.0%	23.6%
	2.5	Count	1062	20	1082
		% within decision_all	33.9%	45.5%	34.1%
	3.5	Count	5	20	25
		% within decision_all	0.2%	45.5%	0.8%
	4.5	Count	0	4	4
		% within decision_all	0.0%	9.1%	0.1%
Total		Count	3129	44	3173
		% within decision_all	100.0%	100.0%	100.0%

star_80 * decision_all Crosstabulation

			decision_all		
			Excluded	Included	Total
star_80	.5	Count	738	0	738
		% within decision_all	46.9%	0.0%	46.3%
	1.5	Count	335	0	335
		% within decision_all	21.3%	0.0%	21.0%
	2.5	Count	498	9	507
		% within decision_all	31.6%	47.4%	31.8%
	3.5	Count	4	8	12
		% within decision_all	0.3%	42.1%	0.8%
	4.5	Count	0	2	2
		% within decision_all	0.0%	10.5%	0.1%
Total		Count	1575	19	1594
		% within decision_all	100.0%	100.0%	100.0%

Forvirringsmatriser:

		Manuell vurdering		
		Inkludert	Ekskludert	
Rayyan	≥ 2.5 stjerner (inkludert)	101	3891	PPV = 101/3992 = 2,53%
	< 2.5 stjerner (ekskludert)	4	2389	NPV = 2389/2393 = 99,83%
		SEN = 101/105 = 96,19%	SPE = 2389/6280 = 38,04%	PR = 105/6385 = 1,64%

Krysstabell: 20% sortert, grenseverdi <2.5 stjerner for ekskludering.

		Manuell vurdering		
		Inkludert	Ekskludert	
Rayyan	≥ 2.5 stjerner (inkludert)	76	2088	PPV = 76/2164 = 3,51%
	< 2.5 stjerner (ekskludert)	1	2631	NPV = 2631/2632 = 99,96%
		SEN = 76/77 = 98,7%	SPE = 2631/4719 = 55,75%	PR = 77/4796 = 1,61%

Krysstabell: 40% sortert, grenseverdi <2.5 stjerner for ekskludering.

		Manuell vurdering		
		Inkludert	Ekskludert	
Rayyan	≥ 2.5 stjerner (inkludert)	44	1067	PPV = 44/1111 = 3,96%
	< 2.5 stjerner (ekskludert)	0	2062	NPV = 2062/2062 = 100%
		SEN = 44/44 = 100%	SPE = 2062/3129 = 65,9%	PR = 44/3173 = 1,39%

Krysstabell: 60% sortert, grenseverdi <2.5 stjerner for ekskludering.

		Manuell vurdering		
		Inkludert	Ekskludert	
Rayyan	≥ 2.5 stjerner (inkludert)	19	502	PPV = 19/521 3,65%
	< 2.5 stjerner (ekskludert)	0	1073	NPV = 1073/173 = 100%
		SEN = 19/19 = 100%	SPE = 1073/1575 = 68,13%	PR = 19/1594 = 1,19%

Krysstabell: 80% sortert, grenseverdi <2.5 stjerner for ekskludering.

		Manuell vurdering		
		Inkludert	Ekskludert	
Rayyan	> 2.5 stjerner (Inkludert)	0	1	PPV = 0/1 = 0%
	≤ 2.5 stjerner (ekskludert)	105	6279	NPV = 6279/6384 = 98,36%
		SEN = 0/105 = 0%	SPE = 6279/6280 = 99,98%	PR = 105/6385 = 1,64%

Krysstabell: 20% sortert, grenseverdi ≤ 2.5 stjerner for ekskludering.

		Manuell vurdering		
		Inkludert	Ekskludert	
Rayyan	> 2.5 stjerner (Inkludert)	14	4	PPV = 14/18 = 77,78%
	≤ 2.5 stjerner (ekskludert)	63	4715	NPV = 4715/4778 = 98,68%
		SEN = 14/77 = 18,18%	SPE = 4715/4719 = 99,92%	PR = 77/4796 = 1,61%

Krysstabell: 40% sortert, grenseverdi ≤ 2.5 stjerner for ekskludering.

		Manuell vurdering		
		Inkludert	Ekskludert	
Rayyan	> 2.5 stjerner (Inkludert)	24	5	PPV = 24/29 = 82,76%
	≤ 2.5 stjerner (ekskludert)	20	3124	NPV = 3124/3144 = 99,36%
		SEN = 24/44 = 54,55%	SPE = 3124/3129 = 99,84%	PR = 44/3173 = 1,39%

Krysstabell: 60% sortert, grenseverdi ≤ 2.5 stjerner for ekskludering.

		Manuell vurdering		
		Inkludert	Ekskludert	
Rayyan	> 2.5 stjerner (Inkludert)	10	4	PPV = 10/14 = 71,43%
	≤ 2.5 stjerner (ekskludert)	9	1571	NPV = 1571/1580 = 99,43%
		SEN = 10/19 = 52,63%	SPE = 1571/1575 = 99,75%	PR = 19/1594 = 1,19%

Krysstabell: 80% sortert, grenseverdi ≤ 2.5 stjerner for ekskludering.