UNIVERSITETET
I OSLO

Masteroppgave

Hvordan skjer sosioteknisk rekonfigurering ved innføring av KI-system i sykehus?

En komparativ studie knyttet til innføring av dyp læring i to norske sykehus

Siv Fjellkårstad

IT og ledelse

60 studiepoeng

Institutt for Informatikk

Det matematisk-naturvitenskapelige fakultet

2023

Hvordan skjer sosioteknisk rekonfigurering ved innføring av KI-system i sykehus?

Siv Fjellkårstad

http://www.duo.uio.no

Trykk: Reprosentralen, Universitetet i Oslo

Sammendrag

Norge står overfor store utfordringer med hensyn til tilgang på helsepersonell, og samfunnet ser seg om etter løsninger. Kunstig intelligens (KI) kan bidra til at oppgaver utføres mer effektivt og nøyaktig, og innenfor radiologi er det nå flere hundre KI-baserte produkter på markedet. Med innføringen av kunstig intelligens, og da kanskje spesielt dyp læring, vil det komme endringer og utfordringer, og det er derfor behov for å studere de som har tatt i bruk dyp læring på sykehus og lære av dem.

Denne komparative studien undersøker erfaringer fra innføring av dyp-læring-systemer (DL-systemer) i to norske sykehus, og svarer på forskningsspørsmålet *"hvordan skjer sosioteknisk rekonfigurering ved innføring av KI-system i sykehus"* ved å ta utgangspunkt i (1) førsituasjonen uten KI-produktet og motivasjon for å ta i bruk kunstig intelligens, (2) viktigste endringer for å komme til den nye konfigurasjonen, og (3) nåværende konfigurasjon med KI-systemet.

Studien er basert på dybdeintervjuer av ansatte som har erfaringer med KI-systemer basert på dyp læring i daglig, pasientnær bruk, og er skrevet for både IT-personell og ansatte i helsetjenesten som vurderer å ta i bruk kunstig intelligens. Jeg har brukt induktiv metode og tatt utgangspunkt i det intervjupersonene har vektlagt som viktig.

Jeg fant at en sentral motivasjonsfaktor for innføringen av DL-systemet i begge casene var ressursutfordringer og forventning om at DL-systemet kunne bøte på disse. Det var vesentlige forskjeller på hvor tett de to sykehusene samarbeidet med leverandøren, leverandørens forretningsmodell og hvordan helsepersonellet brukte KI-systemet på konseptuelt nivå. DL-systemets uløselige forhold til sine treningsdata la føringer for endringsbehov og nødvendige tiltak for å ta DL-systemet i daglig bruk.

Spenning mellom DL-systemets upålitelige natur og bruk på et så sikkerhetskritisk domene som medisin er et område som det er interessant å utforske videre, blant annet for å finne praktisk gjennomførbare løsninger for kvalitetssikring.

Forord

I 1998 sluttet jeg på veterinærhøyskolen for i stedet å ta IT-utdannelse. Jeg fikk høre at de tre årene med veterinærmedisinsk utdannelse var bortkastet – hva nytte hadde man vel av kompetanse på medisin og IT i kombinasjon? I dag går medisin og teknologi hånd i hånd. Varme hender og effektiv teknologi. Denne studien er dedikert til grensesnittet mellom disse to fagområdene, konfigurasjonen mellom helsepersonell og kunstig intelligens, og hvordan denne endres når man tar i bruk kunstig intelligens innenfor radiologi. Jeg håper at både IT-personell og helsepersonell finner glede i, og har nytte av å lese den.

Jeg vil uttrykke min takknemlighet til intervjupersonene i de to sykehusene. Dere er pilarene i studien, ildsjeler og banebrytere på området – tusen takk for tiden jeg fikk av dere.

Jeg vil takke min veileder, Alexander Moltubakk Kempton for tilbakemeldinger, råd og ikke minst stor tålmodighet med alle mine digresjoner og krumspring bort fra den røde tråden i studien.

I tillegg vil jeg takke det norske KI-miljøet i helsetjenesten – alle dere utrolig dyktige ildsjelene tilknyttet sykehus, kommuner, foreninger, clustre, forskningsinstitusjoner og mer. Dere har bidratt til kompetansebygging, engasjement og erfaringsdeling i sektoren, også før ChatGPT ble lansert og kunstig intelligens kom på alles lepper. Jeg har lært så mye av dere, tusen takk for jobben dere gjør!

Oslo, 20/6 2023

Siv Fjellkårstad

Innholdsfortegnelse

Figurer

Tabeller

1 Innledning

1.1 Bakgrunn og motivasjon

Helsetjenesten står overfor flere utfordringer, som økt alder på befolkningen – noe som medfører økt sykdomsbyrde og multimorbiditet[1], større etterspørsel etter helsetjenester, økte forventninger fra samfunnet og økte helseutgifter (Helsepersonellkommisjonen, 2023; Panch et al., 2018). Dette har allerede resultert i merkbart større press på personellet i helse- og omsorgstjenestene, og Norge står, i likhet med andre land, overfor store utfordringer med tilgang på personell. Situasjonen blir enda strammere mot 2040 (Helsepersonellkommisjonen, 2023). Med dette som bakteppe er det ikke unaturlig at samfunnet, inkludert styringsmaktene, ser seg om etter løsninger. Både i Nasjonal helse- og sykehusplan 2020-2023 (2019), *Nasjonal strategi for kunstig intelligens* 2020) og Helsepersonellkommisjonen (2023) pekes det på at bruk av kunstig intelligens (KI) kan bidra til mer bærekraftige helsetjenester, og ifølge *Nasjonal strategi for kunstig intelligens* 2020) vil *"flere oppgaver som i dag utføres av helsepersonell kunne gjennomføres av autonome systemer og kunstig intelligens"*.

Radiologi er et medisinsk fagområde som gjør bruk av bildedannende metoder, som computertomografi (CT), magnetisk resonans (MR), ultralyd (UL) og positronemisjons- tomografi (PET), for å avdekke og behandle sykdommer (Brekke & Borthne, 2022b). Forbedret bildeanalyse, økning i pågang til radiologiske undersøkelser og mangel på radiologer er medvirkende til stor utbredelse av kunstig intelligens innenfor radiologi (Mintz & Brodie, 2019). Antall produkter på området øker. Nettstedet "AI for radiology"[2] inneholdt i 2021 hundre dyp-læring-baserte CE-merkede produkter (DL-systemer) for det europeiske markedet (van Leeuwen et al., 2021). I slutten av januar 2023 var antallet over to hundre (The Radboud university medical center, 2023).

Helsepersonellkommisjonen melder om økende sprik mellom befolkningens forventninger til omfang, økt kvalitet og utbredelse av helsetjenester på den ene siden, og tjenesteyternes mulighet til å møte forventningene som følge av personellmessige og finansielle begrensninger på den andre

[1] Multimorbiditet = har flere sykdommer samtidig. Les mer på https://sml.snl.no/komorbiditet
[2] www.AIforRadiology.com

siden. Forventningene øker med medisinske fremskritt (Helsepersonellkommisjonen, 2023). Trolig har også den enorme oppmerksomheten knyttet til OpenAIs DL-baserte chatbot ChatGPT, som nådde 100 millioner aktive brukere i løpet av to måneder, ført til høyere forventninger til hva kunstig intelligens kan gjøre (Curry, 2023). Gitt forventninger om bruk av kunstig intelligens fra politikere og befolkning, og relativt god tilgang til passende programvare, er det sannsynlig at bruken innen radiologi vil øke i årene fremover. Selv om kunstig intelligens kan bidra til at tjenestene som utføres blir mer nøyaktige og effektive (Rajpurkar et al., 2022), kan innføringen også medføre andre endringer som endret innhold i arbeidsoppgaver og nye oppgaver (Grønsund & Aanestad, 2020), samt en rekke praktiske utfordringer som må håndteres for å få den nye sosiotekniske konfigurasjonen til å fungere (Salwei & Carayon, 2022). Det er derfor interessant å studere de som har tatt i bruk kunstig intelligens i sykehus og lære av dem.

1.2 Problemstilling og avgrensninger

I denne studien undersøker jeg erfaringer fra bruk av kunstig intelligens i spesialisthelsetjenesten. Forskningsspørsmålet er *"Hvordan skjer sosioteknisk rekonfigurering ved innføring av KI-system i sykehus"?*

Jeg bruker begrepet *konfigurasjon*, eller menneske-maskin-konfigurasjon, for å beskrive samspillet mellom mennesker og maskiner. Konfigurasjonen er et sett av relasjoner mellom mennesker og maskiner med en viss fordeling av oppgaver og ansvar mellom dem (Grønsund & Aanestad, 2020). *Rekonfigurering* er prosessen der nye konfigurasjoner av mennesker og maskiner blir til (Mazmanian et al., 2014).

Delproblemstilling 1: Hva var kontekst, motivasjon og mål for innføringen av KI-systemet?

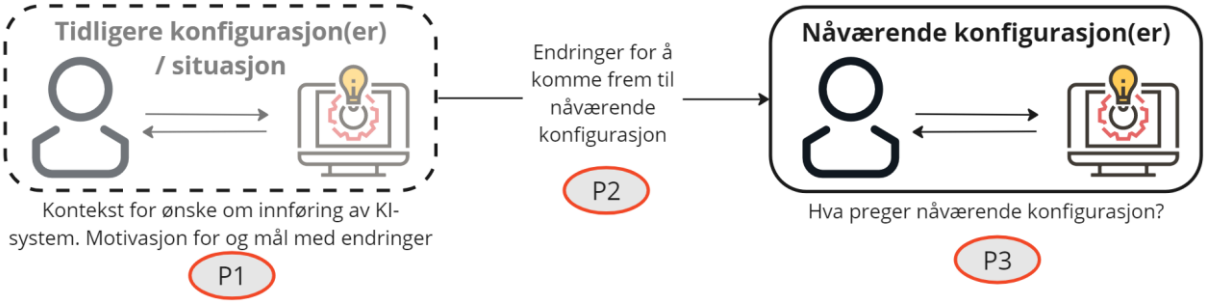
Denne studien fokuserer på en sosioteknisk rekonfigurering, som gjerne skjer når man ønsker å oppnå et mål, som for eksempel å løse en utfordring. Motivasjon og mål med endringene danner derfor basis for den nye konfigurasjonen (Baird & Maruping, 2021), og er nødvendig å kjenne til for å forstå rekonfigurasjonen. I Figur 1 er dette merket med "P1".

Delproblemstilling 2: Endringer for å komme frem til nåværende konfigurasjon

Endringer i prosesser og IT-systemer er tett sammenflettet og styringen går begge veier (Leonardi, 2011). Jeg forventet derfor at når KI-systemet ikke passet helt inn i helsepersonellets arbeidsflyt, så måtte enten IT-systemene (inkludert KI-systemet) eller prosessene endres. Jeg har derfor undersøkt hvilke endringer og tiltak, planlagte eller ikke, intervjupersonene mente var viktige for å ta i bruk KI-systemet. I Figur 1 er dette merket med "P2".

Delproblemstilling 3: Hva særpreger nåværende konfigurasjon?

Siste problemstilling handler om den nye konfigurasjonen. Den kan studeres i flere perspektiver. En deduktiv tilnærming ville vært å undersøke og beskrive den etter et rammeverk, for eksempel Baird and Maruping (2021)s rammeverk for delegering, men dette ville vært for omfattende for denne studien. I stedet ønsket jeg å arbeide induktivt og undersøke hva *intervjupersonene* mente var viktig med den nye konfigurasjonen, for eksempel hva som særpreget den, hvilke oppgaver som var delegert til KI-systemet og hvilket ansvar helsepersonellet beholdt selv. I Figur 1 er dette merket med "P3".



Figur 1 Visualisering av delproblemstillinger

Jeg svarer på forskningsspørsmålet ved å gjennomføre en komparativ, kvalitativ studie basert på to norske sykehus som har tatt i bruk systemer basert på dyp læring innenfor radiologi. I case 1 bruker de DL-systemet som en del av planleggingen av strålebehandling for brystkreftpasienter, og dyp læring brukes for å segmentere strukturer på CT-bilder av bryst. Det vil si at DL-systemet tegner rundt strukturer som enten skal være mål for strålingen ("målvolumet") eller som skal skjermes for strålingen ("risikoorganene"), og som man trenger å kjenne volumet av for å beregne strålingsdose. I case 2 bruker de DL-systemet som en del av diagnostiseringen av pasienter som har kommet til sykehuset med mistanke om fortetninger i arteriene som forsyner hjertet med blod. I dette tilfellet

brukes DL-systemet til å avklare de vanskeligste tilfellene ved å lage en personifisert digital 3D-modell, som deretter brukes for å beregne hvordan blodet flyter gjennom kransarteriene.

1.2.1 Avgrensninger

For å holde omfanget til en håndterbar størrelse begrenset jeg hver case til 4-5 personer som enten hadde erfaringer fra bruk av DL-systemet, førstehånds kjennskap til endringer eller tilpasninger som ble gjort, eller arbeidsoppgaver i nær tilknytning til systemet. Når det gjelder leverandørens DL-tjeneste i case 2, så består den av både menneskelige agenter, og trolig flere maskinelle, hvorav minst en er basert på dyp læring. Jeg ville i utgangspunktet også intervjue leverandørene, men leverandøren i case 1 ønsket ikke å stille til intervju, og siden dette er en komparativ studie valgte jeg da i stedet å gjennomføre litt flere intervjuer av brukerne av DL-systemene. Jeg har derfor holdt det utenfor scope å studere leverandørsiden av DL-tjenesten, men har ved behov sett på teknisk dokumentasjon, videoer og nettsider knyttet til produktet.

Det kan være flere mulige konfigurasjoner med ulike spesifikke arbeidsfordelinger, grader av automatisering og typer oppgavedeling over tid (Mazmanian et al., 2014). I begge casene har de ansatte en arbeidsflyt som totalt sett er bemannet med flere mennesker, maskiner og IT-systemet, men sistnevnte begrenses i denne studien til ett DL-basert system for hvert case. Jeg har også gjort en mindre forenkling og begrenset meg til beskrivelse av *en* sosioteknisk konfigurasjon mellom hvert av de to DL-systemene og de som utfører de samme oppgavene med DL-systemet.

I studien tar jeg utgangspunkt i hvordan DL-modulen (case 1) eller DL-tjenesten (case 2) brukes nå, og hvordan det var før man tok i bruk DL-systemet. Jeg regner rekonfigureringen mellom disse punktene. Mellom dette kan det ha vært flere sosiotekniske konfigurasjoner som jeg ikke har forsøkt å kartlegge.

DL-systemer som har medisinsk hensikt reguleres av regelverket for medisinsk utstyr, og helsetjenester reguleres av flere lover (Helsedirektoratet, 2022)[3]. Disse er viktige rammer for arbeidet med kunstig intelligens, men juridiske problemstillinger er holdt utenfor intervjuguiden for å begrense omfanget av studien.

[3] Merk at jeg har sittet i redaksjonen for kunstig intelligens-sidene på Helsedirektoratets nettsted

1.3 Leseveiledning

Kapittel 1 "Innledning" gjennomgår bakgrunn, motivasjon, forskningsspørsmål med delproblemstillinger og avgrensninger.

Kapittel 2 "Teoretisk bakgrunn" starter med hva kunstig intelligens er, før jeg gjennomgår litteratur knyttet til sosiotekniske systemer, inkludert konfigurasjon, delegering og maskiner som teammedlemmer. Jeg avrunder med litteratur om kunstig intelligens i medisin, spesielt knyttet til nytte, utfordringer og tillit.

Kapittel 3 "Metode" gjennomgår beskrivelse av planlegging av studien, valg av case og intervjupersoner, gjennomføring av datainnsamling, analyse og kvalitet på forskningen. Jeg beskriver her også min rolle knyttet til nasjonal tilrettelegging for KI i helsetjenesten i Norge.

Kapittel 4 "Funn" er delt i en del for hver av de to casene. For hvert case gjennomgår jeg konteksten for ønske om endringer, ulike typer endringer, ny arbeidsflyt med DL-systemet, rammer for bruk, og erfaringer fra innføringen.

Kapittel 5 "Diskusjon" drøfter funnene knyttet til de tre delproblemstillingene sett i lys av litteraturen i kapittel 2. Kapittelet avrundes med implikasjoner for praksis og videre forskning.

Kapittel 6 avrunder studien med en kortfattet konklusjon knyttet til hovedfunn.

Vedleggene omfatter intervjuguide, opplysninger om intervjuene, informasjon om den norske spesialisthelsetjenesten, medisinsk beskrivelse av de to casene og detaljert beskrivelse av behandling av sitater og anonymisering.

1.3.1 Sitater og avidentifisering

Sitater er markert i kursiv og med doble apostrofer, *"slik"*. Ord i doble apostrofer som ikke er i kursiv, "slik", er ikke sitater, men andre ord eller setninger som er naturlig å sette i apostrofer. Øvrig behandling av sitater er beskrevet i vedlegg. Intervjupersonene har uttrykt at det ikke er viktig for dem å være helt anonyme, så jeg har brukt forkortelser for rolle bak alle sitater, siden jeg mener det gir større verdi for leseren av studien. Jeg har både kvinnelige og mannlige intervjupersoner, men har omtalt alle som hankjønn.

1.3.2 Ordforklaringer og forkortelser

Ekstern validering er en evaluering av et systems ytelse på en ekstern kohort som ikke har påvirket utviklingen av systemet (Kleppe et al., 2021).

Implementering brukes i denne studien om å integrere DL-systemet i helsetjenestens systemer slik at det blir en del av de kliniske prosessene. Dette er noe annet enn *utvikling* (= *teknisk* implementering) som er å konstruere DL-systemet ved å lage en modell, trene den og teste den (Makhlysheva et al., 2022). Der det har vært uklarheter om hva som menes, har jeg lagt til "teknisk" eller "klinisk" foran "implementering" for å tydeliggjøre forskjellen.

Trening er optimalisering av modellparametere basert på data (Kleppe et al., 2021).

DL = dyp læring, **ML** = maskinlæring og **KI** = kunstig intelligens**.** Dyp læring er en undergruppe av maskinlæring, som igjen er en undergruppe av kunstig intelligens. Begrepene er beskrevet i avsnitt 2.1. I denne studien bruker til enhver tid det smaleste begrepet som jeg kan gå god for er riktig. Det betyr at jeg ofte har brukt KI der jeg refererer til teori eller rapporter hvor de ikke har brukt et spesifikt begrep, eller hvor jeg er usikker på hvilken teknologi som faktisk er brukt eller studert. Der de spesifiserer at det er snakk om maskinlæring eller dyp læring har jeg brukt det. I funn-delen har jeg brukt DL, siden jeg der vet at produktene bruker dyp læring. Begrepet DL-*system* brukes jeg om et system som bruker en eller flere DL-modeller som en del av beregningene som gjøres (Kleppe et al., 2021).

Medisinsk ordliste ligger i "Vedlegg 4: Medisinsk ordliste".

2 Teoretisk bakgrunn

2.1 Kunstig intelligens

2.1.1 Hva er kunstig intelligens?

Begrepet kunstig intelligens oppstod på midten av 1950-tallet (McCarthy et al., 2006), men det er ennå ingen omforent definisjon av begrepet i litteraturen (Berente et al., 2021; Stokes & Palmer, 2020), og teknologien spenner i dag fra enkle regelbaserte systemer til avanserte dyp-læring-baserte systemer (Sloane & J. Silva, 2020). En definisjon er at KI er et systems evne til å tolke og lære av eksterne data for å oppnå spesifikke mål (Kaplan & Haenlein, 2019), mens en annen er at det er den stadig ekspanderende grensen for hva databehandling kan brukes til (Berente et al., 2021). KI-systemer kan operere autonomt (Baird & Maruping, 2021; Berente et al., 2021) og lære når de får mer data (Baird & Maruping, 2021; Glikson & Woolley, 2020; Lyytinen et al., 2021), men kan også være vanskelige å forstå (Glikson & Woolley, 2020; Lyytinen et al., 2021). Ferdig opptrente systemer er lette å forvirre (Castelvecchi, 2016), for eksempel av statistiske sjeldenheter (Teodorescu et al., 2021).

I mangel på konsensus rundt definisjon av KI i litteraturen, har jeg valgt å bruke en definisjon laget i forbindelse med EUs arbeid med harmoniserte regler for kunstig intelligens (AI Act): "'Kunstig-intelligens-system' betyr et maskinbasert system som er designet for å operere med varierende nivåer av autonomi og som kan, for eksplisitte eller implisitte mål, generere resultater som prediksjoner, anbefalinger eller beslutninger som påvirker fysiske eller virtuelle miljøer" (Bertuzzi, 2023).

Dyp læring, som casene i denne studien er basert på, er en undergruppe av maskinlæring, som igjen er en undergruppe av kunstig intelligens. Den bruker strukturer som etterligner prosessering i den menneskelige hjernen for å gjenkjenne eller identifisere nye mønstre i store datasett (Sloane & J. Silva, 2020). Data kommer inn til systemet via et inndata-lag, og blir deretter evaluert, reprosessert og overført til stadig nye skjulte lag – opptil mer enn hundre, før det endelige resultatet kommer ut av utdata-laget (Holzinger et al., 2019; Shimizu & Nakayama, 2020). Lagene blir kalt skjulte lag fordi

inndata til og utdata fra disse indre lagene ikke er synlige (Mintz & Brodie, 2019). Dette bidrar til at beslutninger basert på dyp læring kan være ekstra vanskelige å forstå (Castelvecchi, 2016).

2.1.2 Forskjeller mellom ML-systemer og tradisjonelle programvaresystemer

Det finnes ulikheter mellom ML-systemer og tradisjonelle informasjonssystemer som kan underminere enkelte av fundamentene for klassiske teorier og konsepter for informasjonssystemer (Teodorescu et al., 2021). ML er først og fremst forskjellig fra tradisjonelle programvaresystemer grunnet den sentrale rollen data har i disse systemene. Til forskjell fra tradisjonelle IT-systemer kan ikke ML-systemer lages uten data, siden systemet blir laget gjennom en avdekkingsprosess basert på treningsdata, og modellen utvikles ved prosessering av data. Dette fører til usikkerheter i de ferdige ML-systemene, og forskjeller mellom ML-systemene og tradisjonelle informasjonssystemer (Ozkaya, 2020).

Å spesifisere et ML-system er i praksis å spesifisere problemet som skal løses, ikke systemet. Siden selve ML-systemet ikke kan spesifiseres og i tillegg inneholder vesentlig mer usikkerheter og "ukjente ukjente" enn tradisjonelle systemer, er utfordringer knyttet til sikker verifisering av systemets fungering kanskje ikke overraskende. En signifikant feilkilde i ML-systemer er også skjulte avhengigheter. Slike finnes i større omfang i ML-systemer enn i tradisjonelle systemer, og gjør det vanskelig å bruke velkjente arkitekturprinsipper for å redusere og kontrollere avhengigheter. Komponentene er både tett koblet og mer komplekse, og prinsippet "endrer man noe, endres alt" (eng. Changing Anything Changes Everything, CACE)" blir derfor brukt om ML-systemer (Ozkaya, 2020).

Dyp læring brukes i dag med hell på en rekke felt (Holzinger et al., 2019). Teknologien har mange fordeler, som at den kan finne mer kompliserte og subtile sammenhenger enn mennesker kan (The Lancet Digital, 2022), og være både rask og nøyaktig i bruk (Rai, 2019), men den har også ulemper. ML-systemer har utfordringer knyttet til begreper som black-box-problemet (Castelvecchi, 2016), ansvarlighet (Martin 2019b), og håndterbarhet (Gunning et al. 2019). I bruk er KI-systemene sårbare for feil, for eksempel ved statistiske sjeldenheter (Teodorescu et al., 2021), og de er dårligere enn mennesker til å forklare hva som ligger bak en beslutning (Barredo Arrieta et al. 2020), spesielt når det er en kompleks modell som ligger bak (Burrell, 2016; Holzinger et al., 2019; Ribeiro et al., 2016). Tiltak som kan føre til bedre forklarbarhet, som robust design (Rosenfeld & Richardson, 2019) og økt

kompetanse hos personellet, er ikke tilstrekkelig for å håndtere utfordringene med kunstig intelligens, siden utfordringene oppstår i KI-systemets trening og påfølgende oppbygging av representasjon av beslutningen, som i liten grad samsvarer med menneskelige semantiske forklaringer (Burrell, 2016).

2.2 Sosiotekniske systemer

For å forstå og beskrive sosiotekniske fenomener trenger man et begrepsapparat. Jeg har valgt å bruke konfigurasjonsbegrepet, og starter med å gjennomgå relevante teorier knyttet til sosiotekniske konfigurasjoner og endringer i disse, *rekonfigurasjon*. Jeg presenterer deretter teori om menneskers og maskiners ulike kapabiliteter, som danner grunnlag for både ønske om å delegere til KI-systemer og innramming av bruken av systemene. Til slutt retter jeg blikket fremover og nevner et par teorier som beskriver mulig retning når teknologien modnes og mennesker kan betrakte KI-systemer mer som likeverdige teammedlemmer.

2.2.1 Sosioteknisk konfigurasjon og rekonfigurasjon

En konfigurasjon er et spesifikt sett av relasjoner mellom menneske(r) og maskin(er) med en gitt deling av oppgaver og ansvar mellom dem (Grønsund & Aanestad, 2020). Interaksjonen mellom informasjonssystemer og mennesker har blitt studert i lys av ulike tradisjoner og perspektiver, og det har vært en diskusjon knyttet til hvorvidt maskiner skal regnes som objekter eller subjekter i en menneske-maskin-konfigurasjon (Suchman, 2007). I starten var litteraturen opptatt av menneskelig styring, hvor menneskenes oppgaver bestod i å bruke teknologiens kapabiliteter, og håndtere begrensningene i teknologien ved å nekte å bruke systemet, endre virksomhetens rutiner eller bruke funksjonaliteten annerledes enn tiltenkt. Etter hvert som det ble lettere å endre teknologien, ble også begrensninger løst ved å lage nye eller endre IT-systemene (Leonardi, 2011). Coiera (2004) hevdet at man i en maskinassistert verden ville oppleve at menneskelig og maskinell styring var så tett sammenvevd, at menneskene som i første omgang *skaper* sosiotekniske systemer i neste omgang *blir formet* av dem (Coiera, 2004), og i dag er det erkjent at styring går begge veier (Leonardi, 2011).

Human-in-the-loop er et begrep som brukes om å supplere kunnskapen som et ML-system får gjennom trening, med menneskets bredere kunnskap. Begrepet brukes både i utvikling, trening (Wu

et al., 2022) og bruk av systemet (Rahwan, 2018). Under bruk av ML-systemer har mennesker en avgjørende rolle i å håndtere utfordrende oppgaver knyttet til overvåkning, unntakshåndtering, optimalisering og vedlikehold, for eksempel ved å identifisere og korrigere feil respons i et ellers autonomt system, eller være en ansvarlig part i tilfelle systemet gir feil respons (Rahwan, 2018), sistnevnte også kalt *human-in-charge* (Kitamura, 2023).

Utvikling i et sosioteknisk system, blir forstått, forklart og navngitt på flere måter. Utviklingen i systemet kan ses på som en serie av stadig nye sosiotekniske konfigurasjoner over tid (Grønsund & Aanestad, 2020), hvor endringene initieres av mennesker, maskiner og andre elementer i konfigurasjonen som kontinuerlig og vekselvis påvirker hverandre til endring (Leonardi, 2011; Mazmanian et al., 2014; Suchman, 2007), og fører til nye mer eller mindre stabile arrangementer (Suchman, 2007). *Rekonfigureringen* kan bestå av endringer i relasjoner, praksiser, teknologidesign og -bruk (Suchman, 2007).

2.2.2 Delegering og rammer for bruk

Allerede da vi fikk de første forsøkene på kunstig intelligens i 50-årene (McCarthy et al., 2006), var man klar over at mennesker og maskiner hadde ulike egenskaper (Baird & Maruping, 2021). Når man tar i bruk KI-systemer kan man derfor spare mennesker for oppgaver som en maskin kan gjøre bedre (Canals & Heukamp, 2019), og mennesker og maskiner kan komplementere hverandre ved å stole på hverandres styrker og overvinne svakhetene (Teodorescu et al., 2021). Mennesker er for eksempel gode på å gjøre vurderinger og valg, mens maskinlæringsmodeller er gode på prediksjon (Teodorescu et al., 2021). Maskiner har god regnekraft og vil kunne håndtere store datamengder både raskt (Rai, 2019; Topol, 2019), pålitelig og med høy presisjon (Rai, 2019). Der mennesker har kognitive bias eller informasjonsmengdene er for store, kan KI-systemene ta bedre beslutninger (Seeber et al., 2020) (Canals & Heukamp, 2019). Imidlertid er KI-systemer fremdeles langt fra gode på forståelse av kontekster, mens mennesker er gode på å umiddelbart skjønne en kontekst og gjøre veldig gode generaliseringer fra få datapunkt (Holzinger et al., 2019), bruke kunnskap i nye kontekster, tenke abstrakt (Véras et al., 2015), utøve kreativitet, empati og gjøre vurderinger (Rai, 2019).

Delegering forekommer typisk når en agent ønsker å frigjøre ressurser for andre oppgaver eller gjennomføre aktiviteter som den ene agenten ikke kan gjøre alene, men for at delegeringen skal skje

må forskjellene være store nok til å rettferdiggjøre kostnadene ved delegeringen (Baird & Maruping, 2021). Ved å for eksempel delegere bort manuelle rutineoppgaver, blir kapasiteten til ansatte frigjort til andre oppgaver (Nedelkoska & Quintini, 2018). Menneskenes ønske om å delegere ansvar til en maskin kan bli til, endres eller forsterkes av en rekke faktorer knyttet til individet, oppgaven eller situasjonen. Aktuelle egenskaper ved situasjonen er *kompleksitet* – hvorvidt situasjonen er stabil, observerbar og kontrollerbar (Rahwan et al., 2019), *stabilitet* – hvorvidt systemet produserer konsistent resultat ved kjente inndata eller kjente lignende situasjoner, *observerbarhet* – hvor transparent systemet er, og *kontrollerbarhet* – i hvilken grad agentene kan kontrollere situasjonen, for eksempel ved å manipulere inndata slik at man får forventet resultat (Baird & Maruping, 2021). Etablering av klare rammer for hvordan kunstig intelligens skal samhandle med omgivelsene gjør at man kan bruke beslutninger fra kunstig intelligens til tross for at det er vanskelig å granske dem, fordi man har god kontroll på agentens virtuelle manøvreringsrom (Asatiani et al., 2021). Dette innebærer å ha god kontroll på treningsdata, data-inn og data-ut, og være oppmerksom i spesifisering av andre rammebetingelser (Robbins, 2020).

Egenskaper ved oppgavene som påvirker delegeringsmulighet er (1) hvilke kognitive, digitale eller fysiske *krav* man har til handlingen (Baird & Maruping, 2021), (2) oppgavens *kompleksitet* – det vil si hvorvidt det er mange avhengigheter, dynamikk og mye usikkerhet knyttet til hvorvidt gitte inndata gir forventede resultater, og (3) i hvilken grad oppgaven kan deles opp, *dekomponeres* (Lee & Siemsen, 2017) (Baird & Maruping, 2021). Individets delegeringsvillighet påvirkes av emosjonelle vurderinger og følelser (Baird & Maruping, 2021), sosiale normer eller ønske om å høste anerkjennelse fra kolleger (Burton et al., 2020), kognitive vurderinger som kostnader, risiko, nytte og hvor lett systemet kan integreres i arbeidsprosessen (Baird & Maruping, 2021), og forventninger til presisjon og systemets kapabiliteter (Burton et al., 2020). Begrepet *kalibrering* brukes for å beskrive hvor godt brukerens tillit til en teknologi stemmer overens med teknologiens kapabiliteter (Asan et al., 2020; Glikson & Woolley, 2020).

2.2.3 Maskiner som teammedlemmer

KI-systemer kan i fremtiden få kapabiliteter som gjør dem til fullverdige og effektive teammedlemmer. En rekke potensielle motsetningsforhold kan gjøre seg gjeldende når man tar i bruk KI-systemer som en kollega. Et eksempel er at man parallelt med økt kvalitet på beslutningene også kan redusere menneskelig evne til å stille kritiske spørsmål. Et annet eksempel er at man kan

oppnå høyere arbeidstempo og bedre kvalitet på arbeidet dersom man bruker et KI-system, men samtidig kan det øke faren for overbelastning av menneskene både fordi de må kvalitetssikre arbeid fra KI-systemer som aldri trenger pause, og fordi man gir mer arbeid til teamet siden man har fått høyere forventning til det (Seeber et al., 2020).

Flere forskere har konseptualisert mulighetene for hvordan man kan få ekstra effekt – positiv eller negativ – av å flette mennesker og maskiner svært tett sammen. Et eksempel er metamenneskesystemer, som er sosiotekniske systemer der mennesker og informasjonssystemer er tett integrert. Gjennom å lære[4] sammen skapes nye kapabiliteter som overgår det mennesker eller maskiner kan oppnå alene, og dette kan føre til betydelige endringer i skala, omfang og hastighet av læring (Lyytinen et al., 2021). Cyber-sosiale systemer er distribuerte nettverk av mennesker og kunstig intelligens. Slike systemer eksisterer allerede i dag, som for eksempel i sosiale media som Facebook, og kan bli utfordrende å håndtere dersom de er uforutsigbare, ikke-transparente, har skjevheter, eller har et innhold hvor det er uklart hva som kommer fra en maskin og hva som kommer fra et menneske. I helsetjenestesammenheng kan man tenke seg cyber-sosiale systemer bestående av store datasett, maskinlæring, og tekniske modeller av klinisk kunnskap (Coiera, 2020).

2.3 Kunstig intelligens i medisin

Denne studien undersøker to caser hvor man bruker dyp læring innenfor radiologi. Jeg vil derfor her gjennomgå bruk av kunstig intelligens på det medisinske fagområdet. Jeg starter overordnet med forventninger, nytte og utfordringer ved bruk av kunstig intelligens i medisin. Jeg går deretter inn i forholdet mellom menneske og KI-system og gjennomgår noe av den omfattende forskningen knyttet til tillit til KI-systemer – med vekt på tillit ved bruk i helsetjenesten, og faktorer som fremmer eller hemmer tillitsforholdet.

2.3.1 Nytte, forventninger og utfordringer ved KI i medisin

De siste årene har klassifisering av bilder med dyp læring blitt stadig bedre. Dette har resultert i en økning i bruk av kunstig intelligens i medisinske fagfelt som lener seg tungt på tolkning av bilder, slik

[4] Forfatterne bruker "maskiner som lærer" for å unngå å begrense begrepet til maskinlæring

som radiologi, patologi, undersøkelser av mage- og tarmsystemet og øyesykdommer. Innenfor radiologi brukes kunstig intelligens i dag innenfor onkologi, kardiologi og nevrologi[5] (Zaharchuk & Davidzon, 2021), og KI-systemer har gitt svært gode resultater på flere medisinske områder, som klassifisering av hudkreft (Esteva et al., 2017), risikostratifisering ved kreft i mage-tarmsystemet (Kleppe et al., 2022) eller diagnostisering av øyesykdommer som diabetisk retinopati (Ting et al., 2017).

Flere studier viser at kliniske eksperter sammen med KI-systemer gir bedre resultater ved for eksempel å være mer presise og konsistente, enn eksperter alene (Rajpurkar et al., 2022; The Lancet Digital, 2022). Andre studier viser at KI-systemer alene på enkelte områder er like bra eller bedre enn medisinsk ekspertise (Holzinger et al., 2020). Innenfor radiologi kan KI for eksempel brukes til segmentering, diagnostisering, og redusere scannetid og strålingsdose (Zaharchuk & Davidzon, 2021). Selv om fordelene med bruk av kunstig intelligens er mange, er helsetjenesten også et komplekst, sikkerhetskritisk domene (Habli et al., 2020). Medisinsk bruk av kunstig intelligens er sårbart, og en feil kan i verste fall føre til direkte skade på mennesker (Habli et al., 2020; Huo et al., 2022). Helsetjenesten har foreløpig liten erfaring med kunstig intelligens, og er i en tidlig fase når det gjelder erfaringer med ekstern validering og klinisk implementering (Rajpurkar et al., 2022). En systematisk gjennomgang av valideringsstudier av DL-algoritmer for bildebasert radiologisk diagnose, viste at det overveldende flertallet rapporterte en redusert ytelse av algoritmen på eksternt datasett, og noen rapporterte en betydelig ytelsesnedgang (Yu et al., 2022). For å sikre at KI-systemet gir forventede gode resultater på sykehusets egne data, er det nødvendig å gjennomføre en valideringsprosess lokalt (Makhlysheva et al., 2022). Medisinske bilder er lettere å validere enn andre datakilder, som naturlig språk, fordi helsepersonellet kan evaluere bildene og dermed direkte validere hvor gode beslutningene fra KI-systemet er (Makhlysheva et al., 2022).

Medisin er av flere grunner likevel regnet for å være et av områdene hvor KI-teknologier har møtt de største utfordringene. En årsak er at informasjonen som brukes i medisinsk beslutningsstøtte har høy usikkerhet, er ufullstendige, heterogene, unøyaktige og kan inneholde feil (Esteva et al., 2017). Det er også vanskelig å definere hva som er grunnleggende sant eller riktig (eng. *ground truth*) i medisin, for eksempel i diagnostisering hvor det mangler en presis forståelse av årsakene til sykdommen (Asan et al., 2020). Kunstig intelligens kan gjøre uventede feil, som er signifikant

[5] Onkologi, kardiologi og neurologi er fagområdene for kreft, hjertesykdom og sykdom i sentralnervesystemet

forskjellig fra feil gjort av mennesker (The Lancet Digital, 2022). Det er også uklart hvordan KI-basert assistanse påvirker menneskelig ytelse. Hvor mye erfaring klinikeren har er av betydning for nytten, og studier har vist at de med mindre erfaring kan ha større nytte av kunstig intelligens (Rajpurkar et al., 2022). Selv om KI-systemer vil fortsette å ta beslutningene på smale områder, vil mennesker fortsatt ha en fordel ved at de kan ta med flere faktorer i beslutninger enn KI-systemet kan, og dermed ta bedre beslutninger enn KI (The Lancet Digital, 2022)

Begge casene i denne studien er på radiologiområdet. KI-systemer kan være til stor nytte for radiologi, og både øke effektiviteten og gi presise diagnoser (Langlotz, 2019). En del artikler har imidlertid advart mot at bruk av kunstig intelligens er en trussel mot radiologer og radiologi, blant annet fordi systemene kan overta oppgaver som radiologene utfører i dag (Obermeyer & Emanuel, 2016; Recht & Bryan, 2017). Imidlertid er det vanligere at mennesket og maskinen deler på ansvaret for en oppgave. Nye oppgaver dukker også opp, noe som sikrer videre behov for menneskelig arbeidskraft (Grønsund & Aanestad, 2020).

2.3.2 Helsepersonellets tillit til KI-systemer

Hvor villige helsepersonellet er til å bruke KI-systemer, har betydning for effekten av at et sykehus skaffer seg et slikt produkt. Holdningen til KI blant helsepersonell varierer, men radiologer er en gruppe helsepersonell som allerede er tekniske av seg, og derfor er generelt positive til kunstig intelligens. For at ansatte i en virksomhet skal akseptere og ta i bruk teknologi, må den imidlertid være lett å bruke, nyttig og innby til tillit (Makhlysheva et al., 2022).

Hvilke faktorer som fører til tillit er forskjellig for roboter, virtuelle systemer og innebygde systemer. For innebygde systemer er tilliten drevet av transparens og pålitelighet, men lav pålitelighet fører ikke alltid til lav tillit og manglende bruk (Glikson & Woolley, 2020). Dårlig transparens trenger heller ikke å ødelegge tilliten, så lenge det finnes evidens på at KI-systemet er godt nok (Makhlysheva et al., 2022). Bedre transparens, robusthet og rettferdighet kan imidlertid øke tilliten til et KI-system (Asan et al., 2020). Oppfattet ekspertisenivå for systemet (Madhavan & Wiegmann, 2007) og type oppgave systemet skal utføre er andre faktorer som spiller en viktig rolle for tillit til systemet (Glikson & Woolley, 2020). Tillit til kunstig intelligens er ikke stabil over tid (Baird & Maruping, 2021). Den kan minske ved dårlige erfaringer (Madhavan & Wiegmann, 2007), og gjentatte positive interaksjoner kan føre til høyere tillit og gradvis overføring av mer rettigheter og ansvar til systemet

(Baird & Maruping, 2021; Ullman & Malle, 2017). I helsetjenesten er det behov for et høyere tillitsnivå enn i enkelte andre sektorer (Rosenfeld & Richardson, 2019). Det er imidlertid lettere å ha tillit til KI-baserte løsninger for å analysere medisinske bilder, siden helsepersonellet kan se bildene før og etter analysen, og har kompetanse til å analysere bildene selv for å kontrollere de KI-baserte konklusjonene (Makhlysheva et al., 2022).

I DL6-systemer er usikkerhet en dominant karakteristika (Ozkaya, 2020), beslutningene kan være vanskelig å forstå (Holzinger et al., 2019), og logikken for å komme frem til et svar er utviklet basert på prosessering av data, ikke ut fra menneskelige spesifikasjoner (Ozkaya, 2020), noe som kan bidra til at feil som eventuelt dukker opp kan komme på uventet tid og være annerledes enn typiske feil begått av mennesker. Dette er situasjonskarakteristika som vi kan gjenkjenne fra risikohåndtering og beredskapsarbeid, og nettopp dette gjør at det kan være relevant å se til litteratur fra pandemihåndteringen. I kjølvannet av Covid-pandemien ble det gjort en undersøkelse i helsetjenesten i USA, knyttet til resiliente organisasjoners evne til å løse uventede problemer med det man har for hånden av verktøy. Tre faktorer var av avgjørende betydning: (1) Evne til å forutse feil, (2) evne til å håndtere feil og tilpasse seg den nye situasjonen, og (3) evne til å gå gjenopprette orden etter feilsituasjonen (Rangachari & L. Woods, 2020). Dette kan ha overføringsverdi til bruk av KI-systemer.

[6] DL = dyp læring

3 Metode

I dette kapittelet forklarer jeg den metodiske tilnærmingen for planlegging og gjennomføring av studien. Til slutt vurderer jeg kvaliteten på studien. Jeg har blant annet brukt metodikklitteratur fra Tjora (2021), Kvale et al. (2015), Okhuysen and Bechky (2009) og Mees-Buss et al. (2022).

3.1 Planlegging av studien

Jeg har dobbel fagbakgrunn med IT-utdannelse og veterinærmedisinske fag, og arbeider i Helsedirektoratet. Dette gjør at jeg har en spesiell interesse for skjæringspunktet mellom medisin og IT, og en indre motivasjon for å bidra til å oversette mellom disse to fagområdene. Jeg ville derfor lage en kontekstrik studie som kunne være interessant og forståelig for både helse- og IT-personell, noe som for eksempel var førende for valg knyttet til tema, filosofisk standpunkt, planlegging av studien og fremstilling av funn.

3.1.1 Valg av forskningsområde

Kunstig intelligens har potensiale til blant annet å forbedre pasientsikkerheten og redusere helsepersonellets arbeidsbyrde, men teknologien er bare ett element av et større sosioteknisk system som må fungere for å få til effektiv bruk (Salwei & Carayon, 2022). Det har imidlertid vært lite forskning på hvordan mennesker og KI-systemer spiller sammen (Grønsund & Aanestad, 2020), og det mangler studier på delegering av ansvar mellom KI-systemer og spesialister i helsetjenesten (Baird & Maruping, 2021). Forskning på kunstig intelligens i virkelige miljøer, som for eksempel organisasjoner som allerede bruker kunstig intelligens-baserte beslutningsstøttesystemer (Glikson & Woolley, 2020), og hvordan disse organisasjonene takler utfordringer relatert til for eksempel forklarbarhet, har i stor grad vært gjennomført på konseptuelt nivå. Det er derfor av interesse å studere erfaringer fra bruk av kunstig intelligens i ekte, komplekse situasjoner i spesialisthelsetjenesten (Asatiani et al., 2021).

3.1.2 Filosofisk tilnærming

En studie basert på dybdeintervjuer vil ha mye fortolkende ved seg. Intervjupersonene er subjektive, og jeg kan ikke unngå å være subjektiv i valg av hva jeg vil fokusere på og hvordan jeg tolker informasjonen jeg får. Subjektiviteten må håndteres for at det som produseres skal kunne få status som forskning, men dette kan gjøres på flere måter. Naturalistisk tilnærming til håndteringen betyr at man bruker strukturerte metoder for å gjøre kvalitative data mer allmenngyldige, men en ulempe ved dette er at den strukturerte prosessen ikke nødvendigvis fører til interessant og plausibel teori (Mees-Buss et al., 2022). Jeg valgte i stedet hermeneutisk tilnærming, som innebærer å følge ledetråder, spor og nye retninger i forskningen, for å få en bedre forståelse for de sosiale fenomenene som studeres. Dette er ifølge Mees-Buss et al. (2022) en grunnleggende basis for fortolkende studier.

3.1.3 Valg av forskningsspørsmål og metode

Første steg i gjennomføringen av en studie er å få tilgang til passende case, *tilgangsforhandlinger* (Pan & Tan, 2011). Jeg startet med en rekke samtaler med mulige informantmiljøer, og basert på disse samtalene justerte jeg forslag til forskningsspørsmål en rekke ganger. Noen elementer i forskningsspørsmålet var imidlertid viktige for meg å ikke rokke ved. Jeg ville at studien skulle være basert på *faktiske erfaringer* med KI-systemer i "skarpe" situasjoner – altså i daglig, pasientnær bruk, siden mye av det jeg hadde lest av artikler tidligere predikerte hva som kom til å skje, i stedet for å undersøke erfaringer fra bruk. Jeg ville også at studien skulle være i skjæringspunktet mellom medisin og IT, siden det er et område jeg har særlig interesse for.

Formøtene tydet på at det var relativt få miljøer som hadde tatt i bruk DL-systemer[7] som en del av helsehjelp i ekte, komplekse situasjoner, og jeg fant heller ikke sosiotekniske studier fra norske miljøer som har DL-systemer i daglig drift[8]. Jeg var usikker på hva jeg ville finne og hvor de

[7] DL = dyp læring. Følgende produkter basert på maskinlæring, f.eks. dyp læring, var i (kjent) bruk i offentlig helsetjeneste i Norge på den tiden: Oncofreeze AI, Intellispace Portal, RayStation, BoneXpert, Paro, HeartFlow, VNNorHurai, RoomMate, AI-Rad Companion, Spider, Syngo.Via og Omsyn. Kilde: https://www.helsedirektoratet.no/tema/KI/Kunstig%20intelligens%20i%20helsetjenesten%20-%202022.pdf (merk at jeg har vært med på å utarbeide rapporten)

[8] Jeg er derimot kjent med en rekke studier gjort i sykehus hvor kunstig intelligens brukes som en del av et forskningsprosjekt

interessante funnene kunne dukke opp, og valgte derfor *kvalitativ metode*, som er egnet for å utforske og forstå et område, inkludert mekanismer og prosesser (Tjora, 2021).

Jeg planla å arbeide induktivt, med mål om å kunne si noe generelt – eller i det minste noe som har gyldighet utover disse to casene, ved å studere enkelttilfeller. Forskningsteori jeg kjenner til påvirket likevel hvordan jeg tolket funnene, så tilnærmingen var kanskje et sted mellom induktiv og abduktiv. Abduktiv tilnærming betyr at man starter induktivt med empirien, men teorier og perspektiver fra forskningsteori spiller inn i forkant av eller i løpet av forskningsprosessen (Tjora, 2021).

For å favne noe av variasjonen i hvordan helsetjenesten har tatt i bruk kunstig intelligens valgte jeg å lage en komparativ studie. En komparativ studie kan være *mellom* eller *innenfor familie* (Bechky BA, 2015). For å øke mulighetene for å få funn med generaliseringsverdi for helsetjenesten valgte jeg et "innenfor familie"-design, hvor "familien" bestod av avdelinger i norske sykehus som hadde tatt i bruk produkter basert på kunstig intelligens i daglig pasientrelatert arbeid, og hadde brukt det lenge nok til å ha fått en del erfaringer. Jeg ekskluderte de som brukte det som en del av et forskningsprosjekt, for å ha så realistiske rammer for bruken som mulig. Begrepet kunstig intelligens favner en stor bredde av teknologier (Sloane & J. Silva, 2020). Nyere generasjoner av kunstig intelligens kan operere autonomt, lære og/eller har redusert granskbarhet (Berente et al., 2021), noe som jeg tenkte kunne påvirke både behov for endringer ved innføring og den nye sosiotekniske konfigurasjonen Jeg ønsket å favne flest mulig av disse fasettene, og valgte derfor å ha krav til at systemene skulle være basert på dyp læring.

3.1.4 Valg av datainnsamlingsmetode

Intervjuer er en egnet innsamlingsmetode når temaet er ulike aspekter av menneskelig erfaring (Kvale et al., 2015; Tjora, 2021), og man ønsker å forstå informantens opplevelser samt hvordan informanten reflekterer over disse (Tjora, 2021). Jeg valgte derfor intervjuer som den viktigste datainnsamlingsmetoden.

I studier av menneskers atferd og interaksjon med omgivelsene, vil observasjoner og uformelle feltstudier vanligvis gi mer gyldig kunnskap enn å intervjue informantene om deres atferd (Kvale et al., 2015). Jeg vurderte derfor feltstudie av bruk av DL-systemene for å studere nåværende konfigurasjon. Jeg gikk bort fra dette både grunnet behov for omfattende søknadsprosess med usikkert utfall, men også fordi jeg var redd for at datainnsamling for en feltstudie av endringer i

konfigurasjon ville måtte gå over relativt lang tid, og derfor gå ut over den tiden jeg hadde til disposisjon.

Jeg vurderte også hvilke punkter i tid jeg skulle samle inn data på for å belyse endringer og tiltak knyttet til innføringen av produktene. Utvikling kan ses på som en serie av stadig nye sosiotekniske konfigurasjoner over tid (Grønsund & Aanestad, 2020). Å studere hvordan disse avløser hverandre frem til nåværende konfigurasjon ville trolig resultert i mer detaljer enn nødvendig for den komparative studien jeg hadde planlagt. Studiens bidrag ville også blitt mer teoretisk og kanskje mindre praktisk anvendbart. I stedet valgte jeg å ta utgangpunkt i to konfigurasjoner for hver case: (1) Den konfigurasjonen i fortiden som de ville endre, og (2) nå-konfigurasjonen hvor de hadde tatt i bruk kunstig intelligens. En praktisk konsekvens av dette er at jeg ikke har med midlertidige eller reverserte endringer, bare de endringene som intervjupersonene nevnte som viktige for nåværende konfigurasjon.

Under intervjuene oppdaget jeg at det var litt ulike oppfatninger av enkelte tekniske temaer og jeg trengte også å vite litt mer om leverandøren for å forstå case 2 godt, så jeg kompletterte intervjuene med informasjon fra leverandørenes nettsider og YouTube-kanaler. Dokumentanalyse utgjorde imidlertid ikke en stor del av datainnsamlingen.

3.2 Valg av case og intervjupersoner

3.2.1 Valg av case

I en komparativ studie må man kunne belyse både fellestrekk og forskjeller (Bechky BA, 2015). Ifølge (Tjora, 2021) blir imidlertid caser ofte ikke valgt optimalt med tanke på dette, men heller pragmatisk på bakgrunn av tilgjengelighet og/eller forskerens kjennskap til dem. Bechky BA (2015) beskrev tre tilnærminger til valg av case: Valg etter kriterier som gjorde at man kunne bygge teori om fellestrekk (eng. *pooled strategy*), valg ut fra store forskjeller i ytelse eller resultater – motpoler (eng. *polar strategy*), og valg av caser med likheter knyttet til et eller enkelte parametre (eng. *matched pair*), for å finne teori som var robust med tanke på kontekstuell variasjon, eller for å finne ukjente årsaker til uventet variasjon (Bechky BA, 2015). Jeg la meg nær sistnevnte og brukte tid på å finne to caser som

både hadde en del likheter som gjorde at man kan sammenligne dem, men samtidig var ulike nok til at jeg tenkte at vi ville få større bredde i funnene og kunne vise til litt variasjon.

Gjennom tidlige samtaler med en rekke miljøer som forsket på eller brukte kunstig intelligens, hørte jeg at det var variasjon i leverandørenes forretningsmodell[9], noe som jeg tenkte kunne være interessant for leseren og også kanskje ville påvirke utviklingen av sosiotekniske konfigurasjoner. Jeg forsøkte derfor å finne to caser som var forskjellige med tanke på leverandørens forretningsmodell. Av tolv miljøer som jeg hadde som mulige caser i begynnelsen, satt jeg igjen med tre miljøer som ut fra dette var godt kvalifisert, hvorav to kunne stille til intervjuer relativt raskt.

Alle miljøene jeg undersøkte var norske, og de to jeg valgte var på norske sykehus.

Case 1 er en enhet for stråleterapi som siden rundt årsskiftet 2021/22 har brukt en dyp-læring-basert modul (heretter kalt "DL-modulen") for segmentering, det vil si at den tegner rundt organer av interesse på CT-bilder, som en del av planlegging av strålebehandling av brystkreftpasienter. Da jeg startet datainnsamlingen var modulen brukt på 30-40 pasienter.

Case 2 er en enhet for angiografi som har brukt en ekstern tjeneste basert på dyp læring siden høsten 2017. Tjenesten brukes for å få ekspertråd i tvilstilfeller knyttet til hvorvidt det er sykdom av betydning i kransarteriene, det vil si blodårene som forsyner hjertet med blod. Også her er det CT-bilder som analyseres. Da jeg startet datainnsamlingen var tjenesten blitt benyttet på et par hundre pasienter.

3.2.2 Valg av intervjupersoner

En god praksis for å velge intervjupersoner er formålbasert utvalg, som betyr å velge informanter ut fra at de har en spesiell erfaring eller ekspertise knyttet til det som skal undersøkes (Johnson et al., 2020), og som av ulike grunner vil kunne uttale seg på en reflektert måte om det aktuelle temaet (Tjora, 2021). Jeg tok derfor kontakt med enhetene, sendte informasjon om studiens innhold og

[9] Dette er oppsummert i kapittel 2.1.1. i https://www.helsedirektoratet.no/tema/KI/Kunstig%20intelligens%20i%20helsetjenesten%20-%202022.pdf (merk at jeg har vært med på utarbeide denne rapporten)

spurte hvem som kunne egne seg som informanter. Jeg fikk tilbake navn på informanter som jeg kunne intervjue.

Intervjupersonene ble plukket ut for å dekke hele arbeidsflyten som var aktuell for studien. For å se om innføringen av DL-systemene også førte til endringer – tilsiktet eller utilsiktet, utover den delen av arbeidsprosessen hvor DL-systemene ble brukt, intervjuet jeg også informanter som arbeidet i prosesstrinnet *før* og *etter* DL-systemet ble brukt, som de som gjennomfører doseplanlegging i case 1, og invasiv[10] kardiolog i case 2. Jeg intervjuet også personale som kjente til de tekniske løsningene i begge casene, for å få overordnet informasjon om teknisk integrasjon, som kan ha noe å si for hvordan samspillet mellom mennesker og teknologi blir.

Valg av informanter er knyttet til en rekke faktorer, som hvilken tilgang man har til mulige informanter, hvor god tid de har, og hvorvidt de er villige til å la seg intervjue (Tjora, 2021). Dette var også gjeldende i mitt tilfelle. Eksempelvis var case 2 et lite miljø, med begrenset antall mulige informanter. Jeg intervjuet fem personer for case 1 og fire for case 2. Jeg vurderte om jeg burde gjennomføre intervju av flere, men det er også fordeler med et lite og relativt spisset utvalg av informanter, som at forskningen blir lettere å håndtere (Brinkmann, 2012). Flere informanter kunne i dette tilfellet redusert gyldigheten, som jeg diskuterer i 3.5.1. Ifølge Kvale et al. (2015) må man imidlertid intervjue "så mange personer som det trengs for å finne ut det du trenger å vite", så da jeg hadde gjennomført de første intervjuene og så hva jeg manglet, spurte jeg om å få lov til å intervjue de som kunne fylle "hullene" i informasjonen som jeg trengte for å svare på forskningsspørsmålet. Underveis i intervjuene fikk jeg også tips om nye intervjupersoner.

3.3 Gjennomføring av datainnsamling

3.3.1 Intervjuer og opplæringsvideo

Intervjuguide

Jeg var opptatt av å få god forståelse for arbeidet til helsepersonellet, konteksten de arbeidet i, og det sosiotekniske perspektivet på bruk av disse produktene. Jeg brukte derfor en vid intervjuguide

[10] *Invasiv* betyr i denne konteksten at man fører instrumenter inn i blodårene for å undersøke dem innenfra

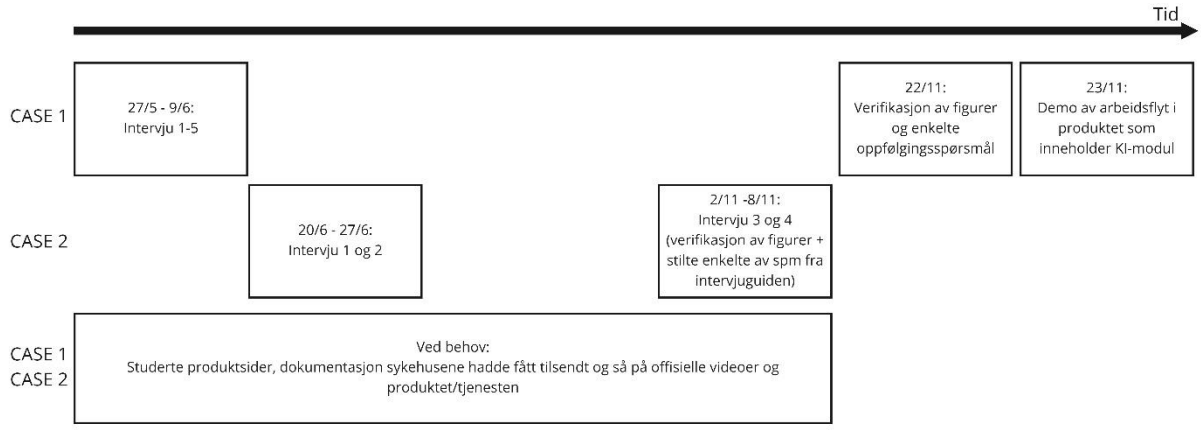
for å belyse problemstillinger i flere perspektiver og lot dem fortelle ganske fritt under hvert spørsmål. Ifølge Tjora (2021) har vi anledning til å bruke det informantene forteller til avgrensning av forskningen etter hvert som den skrider frem, og deltakeren kan bringe inn temaer som ikke nødvendigvis var med i intervjuguiden. Jeg opplevde også å få en del informasjon som jeg ikke direkte spurte etter, og noe av informasjonen som jeg fikk på denne måten var overraskende og gav meg ny og bedre forståelse for temaet. Der det var relevant for forskningsspørsmålet ble derfor disse opplysningene tatt med.

Jeg sendte ut listen over spørsmål i forkant av intervjuene, men ikke alle hadde lest dem på forhånd. Intervjupersonene kunne velge hvilke spørsmål de svarte på, og lengden på intervjuene varierte derfor en del. For detaljerte opplysninger om intervjuene, se vedlegg.

To iterasjoner med videointervjuer

Jeg intervjuet totalt ni personer. Jeg hadde en intervjuguide, og lot personene snakke fritt rundt denne. Jeg stilte oppfølgingsspørsmål når vi kom inn på interessante eller overraskende funn, eller jeg ikke var sikker på om jeg hadde forstått intervjupersonen riktig.

Å være bevisst egne tanker om hva man vil finne er viktig for å senere kunne stille spørsmål ved dem senere. Disse tankene kan bli bekreftet, utfordret eller utvidet etter hvert som man samler inn mer data (Morgan & Nica, 2020). Jeg hadde for eksempel trodd at forretningsmodeller skulle påvirke bruken mer, og hadde derfor stilt spørsmål om kostnader, verdier og samarbeid med leverandør. Noe av dette var for så vidt interessant for problemstillingen, men det var mindre sentralt enn det jeg hadde trodd. Jeg oppdaget også at det var behov for å skjønne arbeidsflyten mye bedre enn jeg hadde trodd i starten, fordi mange endringer var tett knyttet til steg i arbeidsflyten. Jeg trengte derfor å stille flere spørsmål knyttet til denne, tegne den opp og få bekreftet at jeg hadde forstått den riktig. Jeg brukte derfor to iterasjoner på innsamling av data, som vist i Figur 2. Jeg startet med noen intervjuer for hver case, transkriberte og analyserte disse. Deretter skrev jeg ned punkter som jeg trengte å forstå bedre og ba om noen ekstra samtaler for å få svar på disse. For case 1 fikk jeg i tillegg en gjennomgang av en video som viste arbeidsprosessen, laget for opplæringsformål.



Figur 2 Tidslinje for to iterasjoner av datainnsamling

Lydopptak og transkripsjon

På grunn av lange avstander, gjennomførte jeg intervjuene på Teams. Jeg valgte å ta lydopptak av intervjuene og videoopptak av demonstrasjonen, siden transkripsjon direkte i intervjusituasjonen både er distraherende og avbryter samtalens frie flyt. Nøyaktige formuleringer glemmes raskt og intervjueren kan ubevisst filtrere innholdet (Kvale et al., 2015).

Fra samtale via lydopptak og til transkripsjon skjer flere nivåer av abstraksjon hvor man mister kroppsspråk, gester, tempo, stemmeleie og åndedrett (Kvale et al., 2015). For å miste minst mulig gjorde jeg notater underveis i intervjuet, transkriberte fyllord, nøling og pauser, og skrev tidsangivelser i transkripsjonen, for å ved behov kunne gå tilbake til riktig sted i intervjuet.

Muntlig tale er svært forskjellig fra skriftlig, og kan fremstå som usammenhengende og forvirret tale (Kvale et al., 2015). Jeg har derfor kun brukt sitater hvor det var god overensstemmelse mellom ordvalg og det jeg oppfattet som budskapet til intervjupersonen, og gjenfortalt øvrig informasjon.

3.3.2 Offisielle dokumenter og videoer

For å få en god forståelse av produktene DL-systemene var en del av, supplerte jeg intervjuene med offentlig tilgjengelig informasjon fra leverandøren. For case 1 gjaldt dette produktbeskrivelse inkludert noe teknisk beskrivelse. For case 2 fant jeg beskrivelser av tjenesten både på leverandørens offisielle nettsider og leverandørens underleverandør for skytjenester. Jeg brukte også en åpent tilgjengelig video hvor tjenesteleverandøren beskrev hvordan kunstig intelligens ble

brukt i produktet, og hvordan leverandørens analytikerteam arbeidet. Sykehuset sendte meg også dokumenter om tjenesten og om sykehusets integrasjon med tjenesten (skjermbilder og beskrivelse av menyvalg). For beskrivelse av organisatoriske enheter som er intervjuet supplerte jeg intervjuene med informasjon fra sykehusenes offisielle nettsider.

3.4 Analyse

Siden jeg som intervjuer er subjektiv, kan man si at analysen startet med intervjuene – der jeg valgte hvilke utsagn jeg skulle forfølge og stille oppfølgingsspørsmål til og hvilke jeg skulle la passere. Analysen sluttet med skrivingen når jeg bestemte hvordan funnene skulle fremstilles, hva som skulle fremheves og hva som skulle utgå. I det følgende fokuserer jeg på trinnene fra gjennomføring av transkripsjonen.

3.4.1 Empirisk forming av forskningen

Å lytte gjennom intervjuene for å transkribere dem hadde en egenverdi for meg utover at det var nødvendig for transkriberingen. Jeg ble overrasket over at det var flere ting jeg ikke hadde fått med meg da jeg gjennomførte intervjuet, som for eksempel nøling og ordvalg. Ved å lytte til intervjuene på nytt forstod jeg intervjuene bedre. Ting jeg ikke hadde tatt innover meg da jeg gjennomførte intervjuet, falt på plass da jeg transkriberte det.

Jeg la transkripsjonene for hver case inn i en matrise, slik at jeg kunne se alle svarene til en person i sammenheng i kolonnene og svarene på hvert spørsmål på tvers av intervjupersoner i radene. Et skjermbilde av dette er vist i Figur 3. Der intervjupersonene hadde ulike roller og jeg fikk bedre forståelse for hvordan de spilte sammen ved å se dem sammen på denne måten.

Informant 1	Informant 2	Informant 3	Informant 4
3:58: Det var flere: En av dem var at vi så mulighetene for å fjerne repetitive oppgaver som du til	7:20: det er veldig tidsbesparende for oss. Dette med å få strukturer autodefinert kan spare oss for	03:10: Jeg vil si at det er todelt, man kan jo si at det mest opplagte målet, eller motivasjonen, er å redusere	00.20: ja, jeg vil si at motivasjonen var tosidig, og så må jeg bare si noe om bakgrunnen, for det datasettet

Figur 3 Skjermbilde som viser deler av transkripsjonen i analysematrise

Jeg laget en slik tabell for hvert spørsmål. Over hver kolonne er informantnummeret. Klokkeslettene gjør det lettere å navigere i lydfilen.

3.4.2 Analyse av casene

Allerede i transkripsjonstabellen kunne jeg se konturene av svar på enkelte spørsmål knyttet til delproblemstillingene, som konteksten for ønske om endringer, og jeg noterte idéer til bruk i analysen, *empirisk-analytiske referansepunkter* (Tjora, 2021) når de dukket opp.

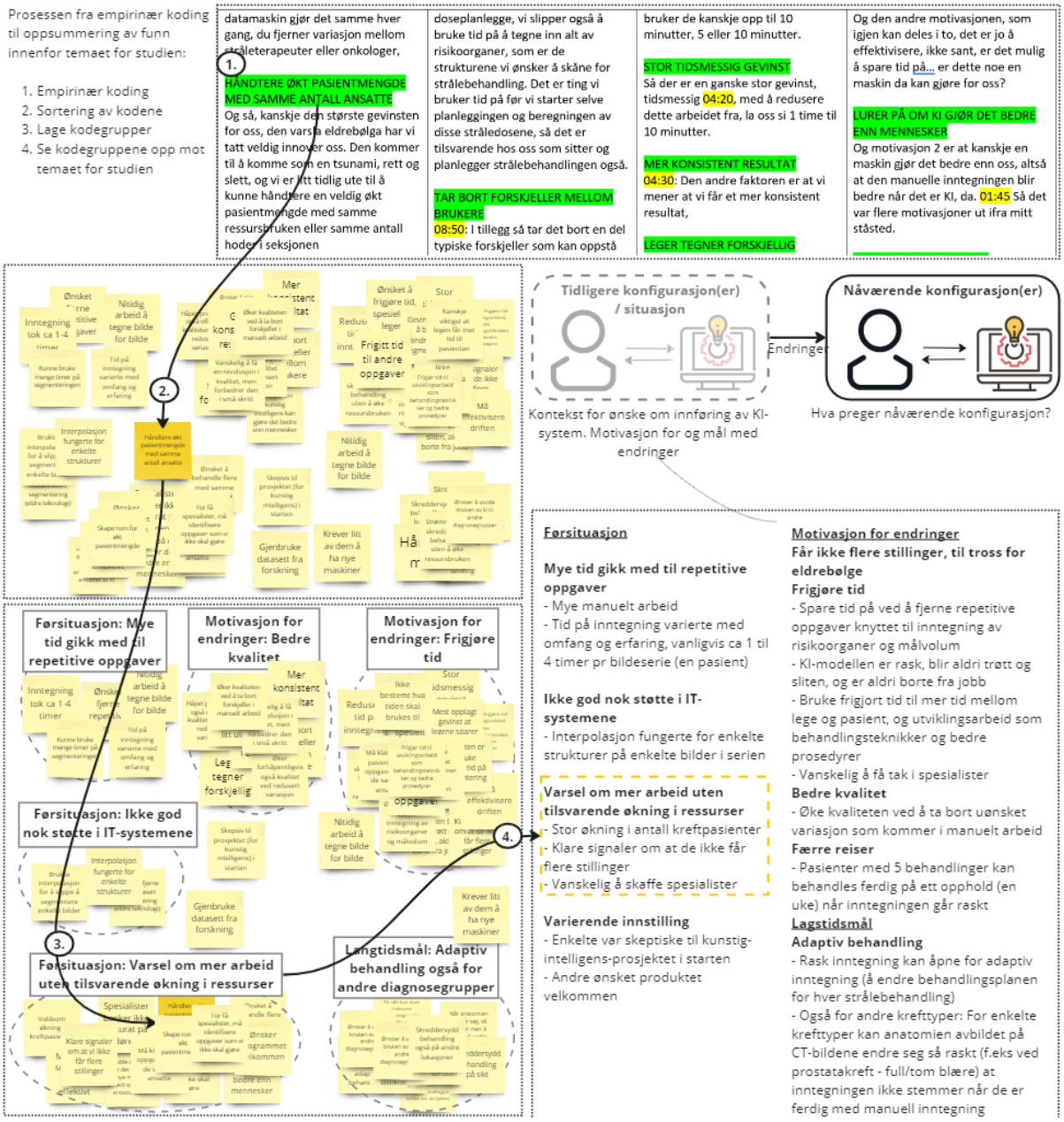
For å gjøre analyse på case-nivå uten altfor mye påvirkning av forventninger og teorier (Tjora, 2021) som jeg kjente til, gjennomførte jeg induktiv empirinær koding. Kodene grupperte jeg sammen tematisk, og hver kodegruppe fikk en tittel. Intervjupersonene hadde snakket relativt fritt, så for å lettere se hva som var relevant for forskningsspørsmålet, og likevel arbeide induktivt, brukte jeg en figur med de tre delproblemstillingene, og satte kodegruppene og litt notater inn i den.

Mennesker og teknologi er så tett sammenvevd at empiriske studier av prosesser eller teknologier egentlig studerer den samme prosessen – den tette sammenvevde styringen som gjøres av mennesker og maskiner sammen (Leonardi, 2011). Jeg studerte derfor arbeidsflyt, DL-systemet og arbeidsflyt i sammenheng. Jeg tegnet opp gammel og ny arbeidsflyt basert på informasjonen jeg hadde fått, og laget varianter med informasjon knyttet til de ulike delproblemstillingene: For å forstå konteksten for ønske om endringer, tegnet jeg den gamle arbeidsflyten og utfordringene knyttet til denne – som utgjorde noe av motivasjonen for endringer. For å forstå samspillet mellom teknologi og mennesker, tegnet jeg inn maskiner som ble brukt, dataflyt, og kvalitetssikringsaktiviteter og andre rammer for bruken av DL-systemet. Jeg studerte informasjonen jeg hadde fra teknisk side, tegnet integrasjoner og arkitekturskisser, og sammenstilte informasjon fra intervjuene med teknisk informasjon som jeg enten fikk tilsendt fra intervjupersonene eller jeg fant liggende åpent på internett. En del av figurene jeg brukte under analysen vurderte jeg som nyttige også for leseren, og disse er derfor lagt inn i kapittelet "Funn".

Jeg brukte den hermeneutiske sirkelen, som betyr at man forstår en kompleks helhet ved å vekselvis forstå helheten, delene helheten består av og deretter helheten igjen (Klein & Myers, 1999), inntil jeg følte at jeg hadde en god forståelse av informasjonen jeg hadde fått. Noen ganger hadde jeg behov for å gå helt tilbake til datainnsamling, og stille oppfølgingsspørsmål eller be om mer informasjon.

Selv om jeg gjorde mye analyse på case-nivå, måtte jeg også beholde komplette og rike data helt frem til sammenligningen av de to casene mot slutten av analysen. Siden det er et element av tid i

forskningsspørsmålet, valgte jeg å også å ha et dokument hvor jeg hadde alle dataene sortert kronologisk under overskrifter som passet empirien, med minimalt av kutt og ingen abstraksjoner.



Figur 4 Skjermbilde av deler av analysen knyttet til induktiv empirinær koding (tidlig i analyseprosessen)

Figur 4 viser et tidlig skjermbilde av deler av analysen fra induktiv empirinær koding til funn innenfor tema for studien. Øverst er transkripsjonen i matrise (som er nærmere beskrevet i Figur 3. Trinn 1 er å legge på empirinær koding for hvert nytt moment som intervjupersonen tar opp. I trinn 2 skrev jeg opp de empirinære kodene på digitale lapper og sorterte dem etter tema. Trinn 3 bestod i å lage kodegrupper og navngi dem. Deretter så jeg kodegruppene opp mot temaet for studien og plasserte dem der de hørte hjemme.

3.4.3 Komparativ analyse og teoretisering

I en komparativ analyse er det ikke bare viktig å finne likheter. Å påpeke variasjon er viktig for å finne nye måter å gjøre ting på og fremme innovasjon på området (Bechky BA, 2015). Den komparative analysen bygget på analysen av de to casene. Jeg brukte tid på å finne riktig abstraksjonsnivå – hvor mye detaljer trengte jeg å abstrahere bort for å finne likheter og tverrgående temaer, og hvor detaljerte måtte dataene være for å finne interessante forskjeller? Dersom jeg for eksempel så helt overordnet på dataene, kunne det virke som at de to sykehusenes tilnærminger til bruk og kvalitetssikring av resultatene fra DL-systemene var helt forskjellig. Dersom jeg gikk inn i detaljene så oppdaget jeg at det var mulig at resultatene fra de to DL-systemene egentlig ble håndtert ganske likt, med helsepersonell som sjekker at resultatet er bra nok før det brukes. Forskjellen var bare at dette ble gjort av sykehusets eget helsepersonell i case 1 og av leverandørens helsepersonell i case 2.

For å gå fra data til konsepter, må man løfte blikket litt og spørre seg hva dette handler om. Finnes det en mer generell "merkelapp" på det vi har funnet i empirien? Finnes det teoretiske bidrag som på en eller annen måte omtaler dette? (Tjora, 2021) Jeg forsøkte derfor å se på funnene på avstand og stille meg disse spørsmålene. I tillegg tenkte jeg på: Hva overrasket meg mest da jeg gjennomførte undersøkelsene? Hva var mest interessant, for meg eller leserne? Hva betyr disse funnene for teorien og praksis? Tanker knyttet til disse spørsmålene ble etter hvert overskrifter i diskusjonskapittelet.

3.4.4 Valg av fremstilling av funn og diskusjon

I diskusjonen presenterte jeg funnene etter delproblemstillingene. Nå ble det klarere hvor jeg trengte stor detaljrikdom og hvor jeg kunne kutte detaljer i funn-kapittelet.

Jeg hadde et ønske om at rapporten skulle være interessant og nyttig for de som arbeider med innføring av kunstig intelligens. I diskusjonen er det nødvendig å se funnene opp mot teorien, og dermed bruke ord og uttrykk fra teorien, noe som kan være fremmedgjørende for enkelte. Jeg

bestemte meg derfor for å presentere *funnene* på en måte som jeg tror er nyttig for både IT-personell og klinisk personell, og bruke begreper fra teorien først i *diskusjonen*.

I tolkningen av et intervju er det av betydning å reflektere over at informasjonen og funnene er sosialt konstruert gjennom interaksjonen mellom forskeren og intervjupersonen (Klein & Myers, 1999) . Funnene er derfor koblet både til hverandre og til konteksten de ble dannet i, noe som gjør at de ikke uten videre er overførbare til en annen virksomhet. For at de skulle forstås riktig forsøkte jeg derfor å gjenskape konteksten for leseren, ved å presentere funnene i kronologisk rekkefølge – i den grad det var mulig, med rik kontekst, språk fra empirien og sortert under overskrifter som er gjenkjennbare for ansatte på sykehus. Jeg la inn enkelte av analysefigurene jeg hadde laget, for å hjelpe leserne med å forstå både funnene og analysen.

3.5 Kvalitet

3.5.1 Gyldighet – relevans og presisjon

Gyldighet knyttes til hvorvidt de svarene jeg har funnet faktisk besvarer forskningsspørsmålet. Gyldigheten kan styrkes ved å tydeliggjøre sammenhengen mellom forskningsspørsmål, valg knyttet til datagenerering og teoretisk grunnlag for analysen. Leseren kan da ta stilling til forskningens relevans og presisjon (Tjora, 2021).

Jeg har redegjort for valg jeg har gjort for å få god sammenheng mellom forskningsspørsmål og valg knyttet til datagenerering og analyse. Her vi jeg derfor spesifikt gjennomgå enkelte valg som påvirker gyldigheten, og som jeg var usikker på underveis.

Intervjupersoner: De to casene var relativt små enheter, og jeg vurderte om antall intervjupersoner var tilstrekkelig. Flere intervjupersoner kunne gitt mer informasjon og større bredde i dataene, men kunne fort også redusert gyldigheten, siden jeg da risikerte å få intervjupersoner som for eksempel hadde mindre eller mer perifer erfaring med DL-systemet. For en komparativ studie, hvor jeg ser hovedtrekk fra casene opp mot hverandre, landet jeg derfor på at antallet intervjupersoner var tilstrekkelig og riktig.

Intervjuguide: I løpet av analysen var jeg flere ganger innom tanken om at det kunne vært interessant å spørre direkte om hvilket ansvar som blir delegert til DL-systemet, siden en viktig del av den sosiotekniske konfigurasjonen handler om delegering av ansvar. Siden fokus i studien er å studere *rekonfigurering* – altså en prosess, ikke nødvendigvis studere den nye sosiotekniske konfigurasjonen (statisk) i detalj, så bestemte jeg meg for at intervjuguiden slik den var dekket problemstillingen.

Intervju- og analysekvalitet: Hermeneutikken innebærer kritisk refleksjon til sosial og historisk bakgrunn knyttet til forskningssituasjonen (kontekstualiseringsprinsippet) (Klein & Myers, 1999), og at man problematiserer utfordringene med fortolkningen, ved stadig å utfordre det deltakerne forteller, og "linsen" som både informantene og intervjueren bruker, samt de teoretiske forklaringene som forskningen resulterer i (Mees-Buss et al., 2022). Jeg tenker at den største svakheten ved intervjuene kan være at jeg har akseptert mye at det som har blitt sagt, i stedet for å utfordre det. Ved å utfordre mer kunne jeg kommet dypere i temaene og for eksempel fått frem utfordringer, som i liten grad ble nevnt av intervjupersonene. Dette fikk også følger for den komparative analysen. Siden jeg utfordret intervjupersonene lite på problemer med innføringen, fikk jeg frem få utfordringer og forskjeller mellom casene.

3.5.2 Pålitelighet – systematikk og transparens

Induktivt rettede kvalitative forskere benytter ikke hypoteser og forsøker ikke å være objektive for å gjøre forskningen repeterbar for andre forskere (Pratt et al., 2020). Refleksjonene som kommer frem fra intervjusituasjonen er *intersubjektive,* avhengige av samspillet mellom deltakeren og forskeren, og analysen bærer også preg av *forskersubjektivitet*. Man kan øke påliteligheten ved å sørge for god sammenheng gjennom hele forskningsprosjektet, med relevante koblinger mellom empiri, analyse og teori, og synliggjøre dette i rapporteringen. Særlig sårbart er utvelgelse av sitater, og hvordan perspektiver eller teorier har bidratt til å inspirere forskningsdesign og analyse (Tjora, 2021). Jeg har derfor redegjort spesielt godt for disse faktorene i metodekapittelet, og redegjort for bruk av sitater i detalj i vedlegg.

Min rolle: Jeg kjente ingen av intervjupersonene på forhånd, men informerte dem om at jeg arbeider i Helsedirektoratet og leder helsemyndighetenes tilrettelegging for bruk av kunstig intelligens i helsetjenesten. I informasjonsbrevet skilte jeg tydelig på rollen jeg hadde overfor

intervjupersonene i studien, og den jeg har i profesjonell sammenheng. Min rolle er likevel ikke helt uvesentlig for reliabiliteten, siden man kan tenke seg at de justerer uttalelsene sine basert på forestillinger av hvem jeg er, hvor jeg arbeidet og hva de tror jeg mener. I tillegg har jeg gjennom arbeidet mitt fått god innsikt i helsetjenestens arbeid med kunstig intelligens, og denne innsikten kan ha påvirket analysen, kanskje helst på slutten når man går fra empiri til konsepter og teori.

Tydelig på hvem som sier hva: I kvalitativ forskning er det viktig å ikke bare fortelle leseren, men også vise leseren (Golden-Biddle & Locke, 2007). Jeg har derfor brukt mye sitater og ellers gjenfortalt innholdet så presist som mulig i Funn-delen. Siden det kan være praktisk for leseren å vite hvem som har sagt hva (Tjora, 2021), har jeg skrevet hvilken informant hvert sitat eller påstand kommer fra. Der jeg selv har tolket eller analysert noe, har jeg eksplisitt skrevet det.

Intervjuerens egenart og bakgrunn: Jeg har beskrevet fremgangsmåten jeg har brukt i detalj, og tror at andre som følger samme fremgangsmåten ville kommet til ca det samme som meg, men jeg tror også at resultatene kunne variere noe som en følge av at intervjuere er forskjellige. Dersom noen for eksempel bruker samme intervjuguide, men utfordrer mer under intervjuet og stiller litt andre oppfølgingsspørsmål, så kan de ende med litt andre funn og konsepter på slutten av arbeidet. De vil også trolig ha en annen bakgrunn enn meg og derfor tolke informasjonen annerledes.

3.5.3 Generaliserbarhet

Diskusjonen om nødvendigheten av generaliserbarhet i kvantitativ forskning, og hvordan det i så fall skal gjøres, har pågått i lang tid (Tjora, 2021). *Moderat generalisering* sier at generalisering ikke kan overlates til leseren, så forskeren må forklare hvilke situasjoner, kontekster, tider og steder forskningen kan generaliseres til (Payne, 2007). Kvale et al. (2015) bruker begrepet *analytisk generalisering*, og påpeker at man må vurdere hvorvidt et funn i en situasjon kan brukes som indikasjon på hva som vil skje i en annen situasjon.

Jeg har brukt *konseptuell generalisering*, som er når utviklede konsepter, typologier eller teorier har relevans for andre tilfeller enn de som er studert. Disse bygges opp empirisk fra empirinære koder, via kodegrupper, til konsepter som ses i sammenheng med teori. Konseptene, for eksempel modeller, begreper og metaforer, er ikke spesifikt knyttet til empirien. Man sikrer relevans for andre ved å se dem opp mot tidligere forskning og teorier, og på den måten gi dem større gyldighet og

generaliserbarhet (Tjora, 2021). Metaforene "arbeidsvillig assistent" og "kompetent kollega" er eksempler på konseptuell generalisering.

4 Funn

4.1 Case 1 DL-basert autosegmentering

Enhet 1 er en enhet for stråleterapi som fra årsskiftet 2021/22 har brukt en dyp-læring-basert modul (heretter kalt "DL-modulen") for segmentering ved planlegging av strålebehandling for brystkreftpasienter. Den var ifølge stråleterapeuten brukt på 30-40 pasienter da intervjuene ble gjennomført. Doseplanleggingen blir gjort som et samarbeid mellom lege, stråleterapeut og medisinsk fysiker. Jeg intervjuet følgende roller: Seksjonsleder (sl), stråleterapeut (st), medisinsk fysiker (f), lege i spesialisering (l) og overlege/onkolog (o).

Etter at en pasient er operert for brystkreft, får vedkommende strålebehandling. En del av planleggingen av strålebehandlingen er ta CT-bilder av området og tegne rundt ("segmentere") strukturer som enten skal være mål for strålingen ("målvolumet") eller som skal skjermes for strålingen ("risikoorganene"). Denne inntegningen ble automatisert da sykehuset tok i bruk DL-modulen.

DL-modulen eies av et skandinavisk selskap og brukes av sykehus verden over, hovedsakelig i USA, Europa og Asia. Ifølge informasjon på produsentens nettside fra mars 2020, tilbyr selskapet kundene å bruke ferdigtrente modeller fra ledende kreftklinikker, men kundene kan også trene modellen med egne data. Modulen er en del av et produkt sykehuset allerede bruker, og to norske sykehus har samarbeidet med leverandøren om å lage testdata i form av CT-bilder med nøyaktig inntegning av strukturer av interesse. Nøyaktig inntegning er viktig for å lage en optimal stråleplan, for "du ønsker å gi så høy dose som du kan til det området [som er sykt], og så spare i mest mulig grad [friskt vev]" (sl). Hensikten med KI-modulen er å gjøre denne inntegningen raskt og effektivt.

4.1.1 Konteksten for ønske om endringer

Signaler om at man ikke kunne ansette mer personell til tross for at eldrebølgen medfører flere pasienter, var en viktig årsak til arbeidet med kunstig intelligens på Sykehus 1: *"Den kommer til å komme som en tsunami, og vi er litt tidlig ute til å kunne håndtere en veldig økt pasientmengde med samme ressursbruken eller samme antall hoder i seksjonen" (sl). "Det er en voldsom økning i antall*

kreftpasienter og (…) vi må identifisere arbeidsoppgaver som vi ikke skal gjøre, for det er ikke slik at det står mange spesialister i onkologi og banker på døra" (o).

Manuell segmentering er tidkrevende og pirkete arbeid. En vanlig CT-undersøkelse av brystregionen utgjør 150 bilder[11], og på hvert bilde skal det tegnes inn et variabelt antall strukturer. Fysikeren anslo tidsbruken til *"typisk en time for en erfaren onkolog",* og leger i spesialisering (LIS), som har lite erfaring, kunne bruke helt opp til en dag. De ønsket å *"arbeide med andre ting enn å sitte med det nitidige arbeidet det er å tegne bilde for bilde" (st)*. De håpet at bruk av KI ikke bare kunne frigjøre tid, men også øke kvaliteten på segmenteringen. *"Vi definerer hva som skal strålebehandles litt ulikt, det er litt forskjell mellom oss (…) Kunstig intelligens gjør at det er mer enhetlig hvordan målevolumet defineres, så forhåpentligvis bedrer det også kvaliteten" (l)*. Gevinsten var *"ikke den rene tidsbesparende effekten" (st)*, men det de kunne *bruke* tiden til: Å øke kapasiteten, og bruke mer tid til dialog med pasienten, utviklingsarbeid som forbedret behandlingsteknikker og bedre prosedyrer, og opplæring. Tabell 1 Case 1: Ønsket bruk av spart tid

oppsummerer mål som intervjupersonene nevnte knyttet til det å spare tid.

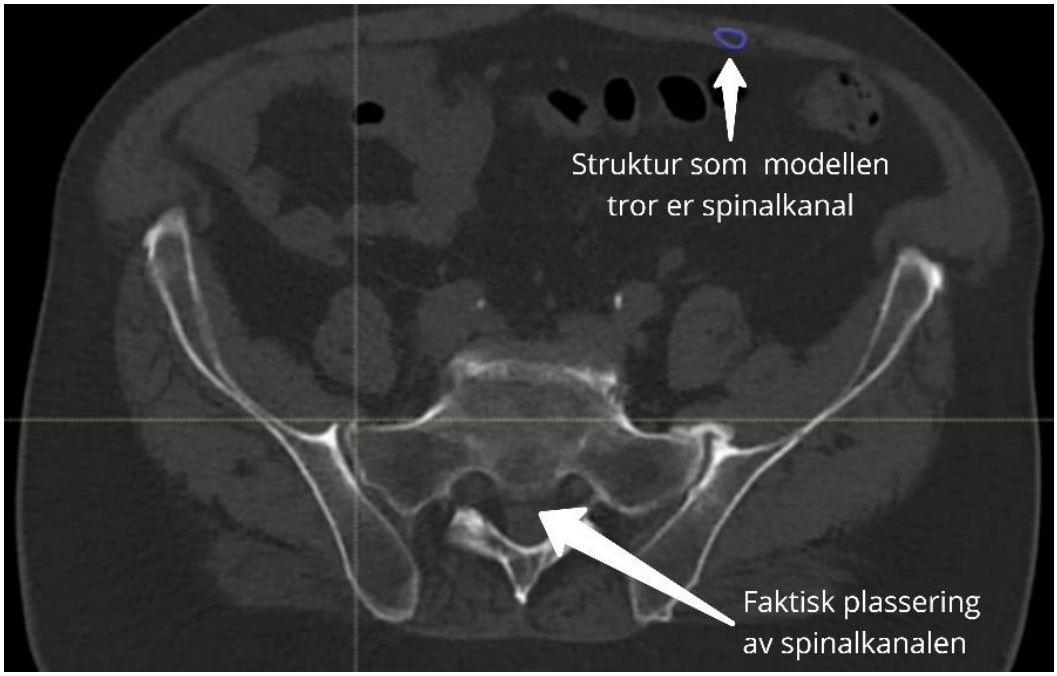
Bruk av spart tid	**Beskrivelse/sitat**
Øke kapasiteten	"Jeg tenker at vi må skape rom for den økte pasientmengden som kommer" og "vi må identifisere arbeidsoppgaver som vi ikke skal gjøre" (o)
Bruke større andel av tiden på pasienten	"Det betyr at legen i stedet for å sitte bak en PC (…) får mer tid med pasienten, enn bak dataskjermen, og det er stor pasientverdi" (st)
Spare pasientene for reiser	Raskere segmentering kan spare pasienter for ekstra turer, fordi mer av arbeidet kan gjøres på ett sykehusbesøk
Mer tid til utviklingsarbeid og opplæring	• Rom for å "drive utviklingsarbeid, som gir en del sekundære gevinster. Man kan utvikle nye behandlingsteknikker, man får bedre prosedyrer

[11] Ca-tall ved 3 mm avstand mellom bildene

	fordi man kan arbeide med andre ting enn å sitte med det nitidige arbeidet det er å regne bilde for bilde" (st) • "*fagutvikling"* og *"bedre utdanning av nye spesialister*" (o)

Tabell 1 Case 1: Ønsket bruk av spart tid

Før innføringen av DL-modulen ble inntegningen gjort manuelt med digitalt tegneverktøy og støtte av en interpoleringsfunksjon[12]. I noen tilfeller kunne man bruke modellbasert autosegmentering (ikke DL-basert), og *"for noen strukturer fungerer den ok (…) mens den ville vært helt sjanseløs for eksempel på et spiserør, som har en ganske diskret framstilling i forhold til omgivelsene rundt, [og] så er det ganske likt hjertet i tetthet"* (st). *"Så erfaringene våre er at vi ofte bruker like lang tid på å korrigere en slik struktur som vi bruker på å tegne den"* (st). Figur 5 viser eksempel på feilsegmentering med denne teknologien.



Figur 5 Case 1: Skjermbilde av feil ved modellbasert autosegmentering (ikke kunstig intelligens)

[12] Interpoleringsfunksjonen gjør at de for eksempel kan tegne på annethvert bilde, og få forslag til inntegning på de andre

Merk at modellbasert autosegmentering i dette tilfellet tar svært feil av plassering av spinalkanalen. Helsepersonellet må da flytte den blå segmenteringen til der den burde vært og se om programmet da kan segmentere spinalkanalen på riktig sted.

Helsepersonellet så på arbeidet med DL-modulen som et skritt mot større gevinster, og de hadde planer om forbedring og nyutvikling i flere retninger. Tabell 2 Case 1: Langsiktige mål med innføringen av DL-modulen oppsummerer langsiktige mål nevnt av intervjupersonene.

Langsiktig mål	**Beskrivelse**
Adaptiv strålebehandling	Ett mål er å *"gjøre en såkalt adaptiv, eller skreddersydd, strålebehandling i fremtiden"* (f). Det vil si at man tar nye CT-bilder hver gang pasienten skal ha behandling, slik at bildene passer helt med *"dagens anatomi"*[13] hver gang. *"Da kan man i prinsippet gi strålebehandling som enten har mindre bivirkninger eller høyere effekt enn i dag"*, men *"i dag kan det være litt utfordrende hvis man trenger å bruke veldig lang tid på å tegne inn slike strukturer på nytt"* (f)
Videreutvikling for bruk på andre bilder fra samme anatomiske region	"*På en lungekreftpasient tegner vi også inn hjerte, vi tegner inn lunger, spiserør, pusterør, spinalkanal, og da kan vi bruke modellen til de også, for å få autogenerert disse strukturene*" (st)
Utvikle nye modeller for andre pasientgrupper	Vi *"er godt i gang med nye modeller allerede, i andre [kropps]regioner"*[14] (sl)
Utvikle modeller som gir resultat man ikke trenger å justere	*"Så går det jo an å sende det datamaterialet tilbake til leverandør som igjen trener modellen, så den modellen etter hvert blir enda bedre. Kanskje den kan – etter hvert – bli enda bedre og kan klare å ta ut de siste*

[13] Her menes det at man kan tilpasse behandlingen til hvordan organene ligger ved hver behandling. Plassering vil være litt forskjellig fra behandling til behandling. Forskjellene vil være større ved enkelte andre kreftformer, f.eks. prostatakreft, som ligger nær urinblæren

[14] Kroppslige regioner

	prosentene også, så det ikke blir behov for noe manuell justering i fremtiden" (st)

Tabell 2 Case 1: Langsiktige mål med innføringen av DL-modulen

4.1.2 Preimplementering og implementering

Preimplementeringsfasen og implementeringsfasen i case 1 gikk ut på å lage og kvalitetssikre treningsdata, og gjennomføre en validering av modellen før KI-modulen ble tatt i bruk. Her gjennomgår jeg disse oppgavene.

Treningsdata og validering

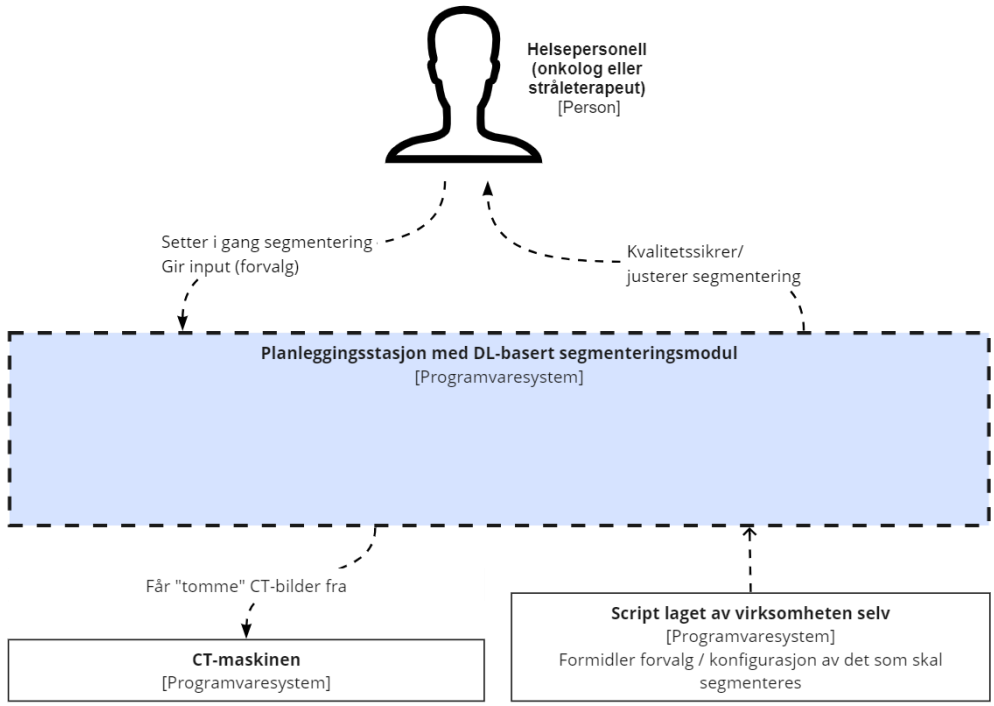
Sykehuset hadde i samarbeid med et annet norsk sykehus segmentert over 200 CT-bildeserier på ca 150 bilder hver, for bruk som treningsdata. *"Det er ganske mange timer som skal til for å samle inn datagrunnlaget"*, forteller stråleterapeuten, og *"i arbeidet med modellen så blir det gjort endringer i forhold til å prioritere tid. Vi har jo ikke stoppet produksjonen, så alt dette er jo arbeid (..) i tillegg til vanlig drift"*. For å få godt treningsmateriale som mulig, brukte de *"ESTRO Guidelines, den europeiske stråleorganisasjonens guidelines for målvolumsdefinisjon for tidlig brystkreft"* (o) for segmentering, og mesteparten av dataene ble generert på de samme bildedannende enhetene som de bruker i det daglige. *"Da vi laget produktet var vi tre overleger som tegnet alle disse CT-ene, og da vi gjorde det første gangen, og sendte det inn til <leverandør>, så vi at det var litt forskjell mellom oss (…) så vi måtte sette oss sammen og harmonisere litt for at det skulle bli et godt produkt"* (o). I tillegg laget de script for *"å konfigurere pasientdataene (…) og sørge for at dataene er ryddige og systematiske og har et format som er egnet til å lære opp en modell"* (f). De samarbeidet tett med leverandøren: *"Så vi gjør arbeidet [lager treningsdata], og så sender vi alt av data til dem, og så trener de modellen og sender den tilbake til oss til validering. Når vi har validert den, så implementerer de det [teknisk] i sitt system"* (sl). Om valideringen fortalte fysikeren at *"vi gjennomførte en validering av denne modellen i fjor høst, og vi testet denne modellen på 15 pasienter[15] og vi så at den produserte et resultat vi var fornøyde med"* (f).

[15] 15 pasienter betyr i praksis' 15 bildeserier på ca 150 bilder

Teknisk integrasjon og leverandørsamarbeid

Det var lite arbeid knyttet til teknisk integrering av KI-modulen, "det er egentlig ingen ting nytt ennå med KI, softwaren hadde vi, hardwaren hadde vi, strålemaskinene var der, og alle systemene rundt var på plass. Da vi fikk lisensen til kunstig intelligens, så ble den bare lagt til (…) Så ingen ting er endret utover at vi fikk tilgang til en modul til" (s). Modulen kunne brukes ved å trykke på en knapp, men "så har vi valgt selv å ta den integrasjonen enda et skritt videre, ved å gjøre slik at våre script automatisk kjører den modellen og tegner disse strukturene. Så vi har integrert enda litt bedre enn det som kommer ut av pakken i programmet" (f).

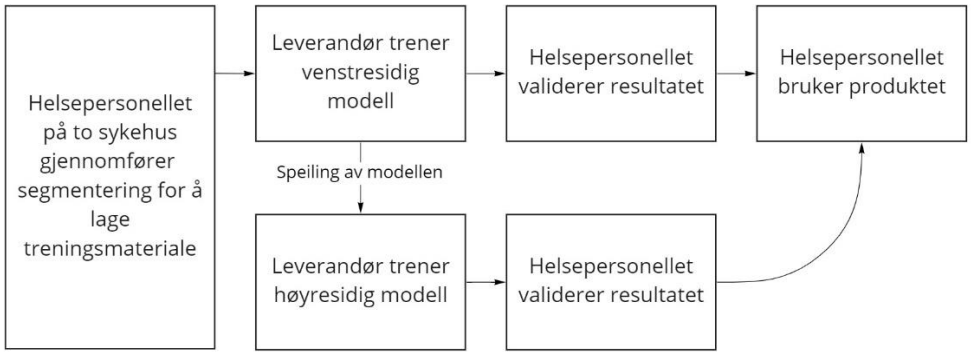
Figur 6 er en forenklet visualisering av konteksten for DL-modulen. Segmenteringsmodulen er plassert i midten og markert med lys blå bakgrunn. Over modulen er brukerne av modulen plassert. Under er systemer som modulen får inndata fra.



Figur 6 Case 1: Systemkonteksten for DL-modulen

Pilene viser prosessflyt, og skal leses i pilens retning, eksempelvis: "helsepersonell" + "setter i gang segmentering" på "planleggingsstasjon med DL-basert segmenteringsmodul".

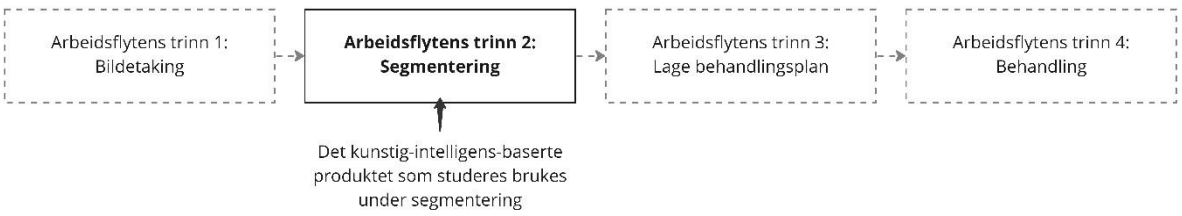
Fysikeren fortalte at de har videreutviklet produktet: "Dette prosjektet var jo opprinnelig for venstresidig brystkreft, og så har vi i etterkant kjørt et prosjekt for høyresidig brystkreft", og "leverandøren klarte rett og slett å speile denne modellen". Også etter dette måtte de kjøre en valideringsjobb. Figur 7 viser arbeidet som ble gjort i forbindelse med denne høyresidige modellen. De fikk bekreftet at det fungerte tilfredsstillende, selv om denne modellen ikke har like gode resultater som modellen for venstresidig brystkreft, "for der er ikke støttestrukturene tegnet manuelt (…) man har speilet strukturene på venstre side over på høyre, sånn at for pasienter med høyresidig brystkreft er det som regel mer korreksjonsarbeid enn for de venstresidig kreftpasientene. Det er litt mer avvik på strukturene der" (st).



Figur 7 Case 1: Beskrivelse av samarbeid i utviklingsprosess for de to DL-modellene

4.1.3 Arbeidsflyt med DL-modulen

Prosessen for strålebehandling kan grovt deles opp i fire trinn, som vist i Figur 8: (1) Ta CT-bilder (2) segmentere bildene (3) lage behandlingsplan og (4) gi strålebehandling. Produktet som studeres brukes under segmenteringen, og vil beskrives detaljert i det følgende. Jeg vil også gi en overordnet gjennomgang av arbeidsflyten og de viktigste beslutningene som tas før og etter dette trinnet, der det bidrar til å gi kontekst og belyse forskningsspørsmålet.



Figur 8 Case 1: Hovedtrinn i arbeidsflyt for postoperative brystkreftpasienter

Forarbeid og billedtaking

Den første delen av arbeidsprosessen, bildetaking, er relevant fordi det blir gjort handlinger der som påvirker CT-bildene som skal analyseres. Onkologen fortalte at *"når jeg har pasient på poliklinikken, så snakker jeg med den pasienten, og så går jeg gjennom histologisvar, så går jeg gjennom bilder, og så vurderer jeg hva pasienten skal ha strålebehandling mot*[16]*"*. LIS-legen fortalte videre at de legger blytråder rundt det aktuelle brystet på pasienten *"akkurat rundt der vi klinisk ser at brystet går (…) Det er for at det skal bli lettere for både oss og AI*[17]*-systemet å finne konturen av brystet [på CT-bildene] "*. Deretter tok man CT-bilder av kvinnenes brystregion i samme stilling som hun skulle ligge under behandling.

Segmentering

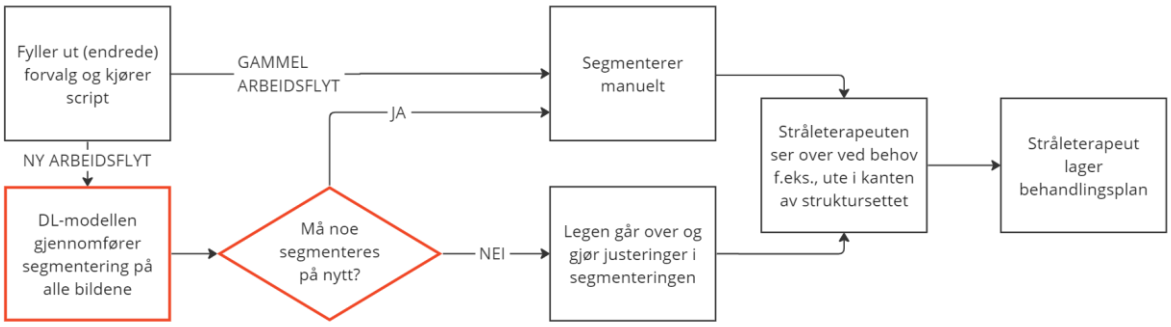
Det er viktig at man treffer godt med strålene, så det som skal behandles får en stor nok dose samtidig som at organene rundt får så lite som mulig. Trinn 2, segmentering, er et ledd i å sikre dette, hvor CT-bildene blir kjørt gjennom DL-modulen som streker rundt grensene for strukturer som skal ha stråling (målvolumer), eller skånes for stråling (risikoorganene), til sammen ca 30 strukturer for hver pasient. Før DL-modulen kan sette i gang med dette, trenger den input fra helsepersonellet på ulike forvalg, som hvor store marginer som skal legges på når man tegner rundt strukturene.

CT-bildene blir kjørt gjennom DL-modulen som identifiserer og streker rundt målvolumene og risikoorganene. Figur 9 viser hovedforskjellen på gammel og ny arbeidsflyt: I dag er det DL-modulen som gjennomfører segmentering på alle bildene i denne pasientgruppen, mens helsepersonellet har fått en ny oppgave knyttet til å beslutte hvorvidt hver del av arbeidet gjort av DL-modulen skal brukes eller segmenteres på nytt. Kvalitetssikringen av strukturene er likevel en liten jobb i forhold til den manuelle segmenteringen som de gjorde før. Stråleterapeutene og onkologene deler på arbeidet med kvalitetssikring. Onkologen fortalte at *"hvis det er jeg som skal lage stråleplanen, så setter jeg meg ned ved tegnebordet, og tar opp denne [ferdig genererte] planen, og så går jeg gjennom snitt for snitt og bare kontrollerer at jeg er fornøyd med segmenteringen. Da går jeg gjennom både risikoorganene, i hvert fall de risikoorganene jeg mener er viktige, og det er jo hjerte*

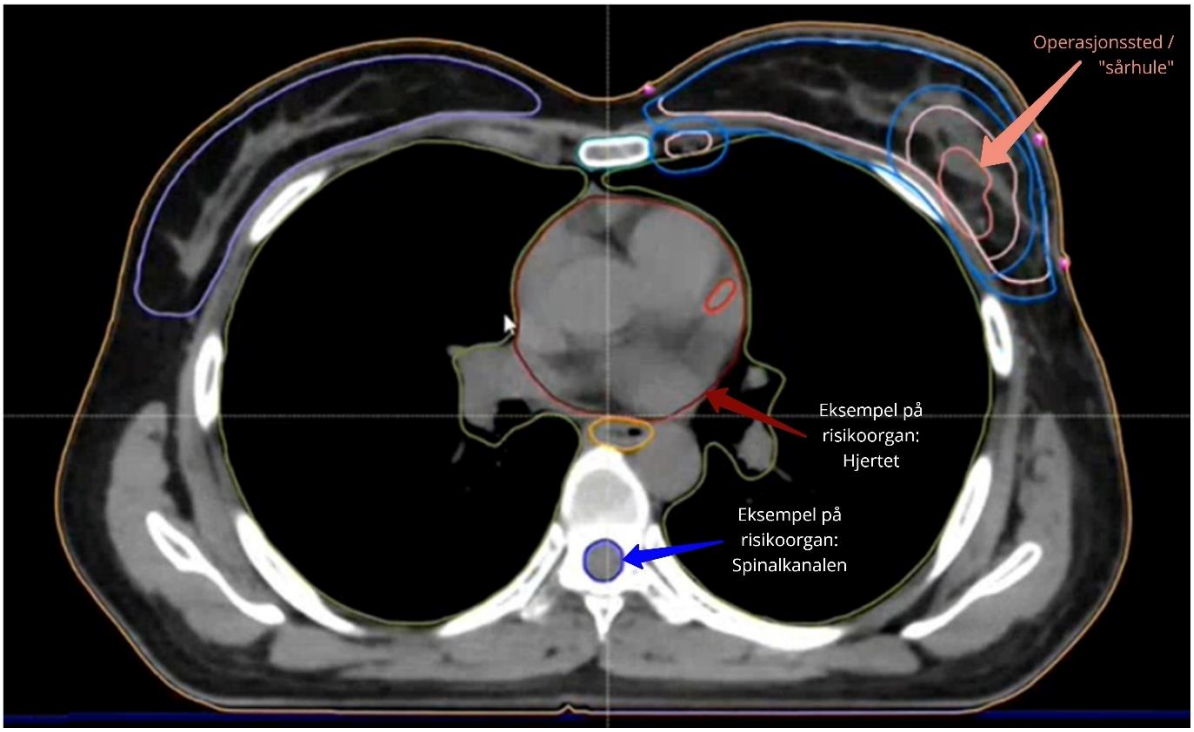
[16] Hvilke målvolumer som skal strålebehandles

[17] AI = Artificial Intelligence (eng.), her menes altså den DL-baserte modulen

og lunge, thyroidea[18] og esophagus[19], og så går jeg også gjennom det jeg ønsker å strålebehandle, det vi kaller målvolum".



Figur 9 Case 1: Overordnet sammenligning av gammel og ny arbeidsflyt (for erfarne leger)



Figur 10 Case 1: Skjermbilde av segmentering av brystregionen utført av DL-modulen

Bildet er et tverrsnitt gjennom kroppen, med ryggsøylen nederst og brystene øverst. Ulike strukturer er markert med forskjellig fargekode for å være lette å gjenkjenne. Området som skal få strålebehandling (brystet oppe til høyre) er markert

[18] Thyroidea = skjoldbruskkjertelen

[19] Esophagus = spiserøret

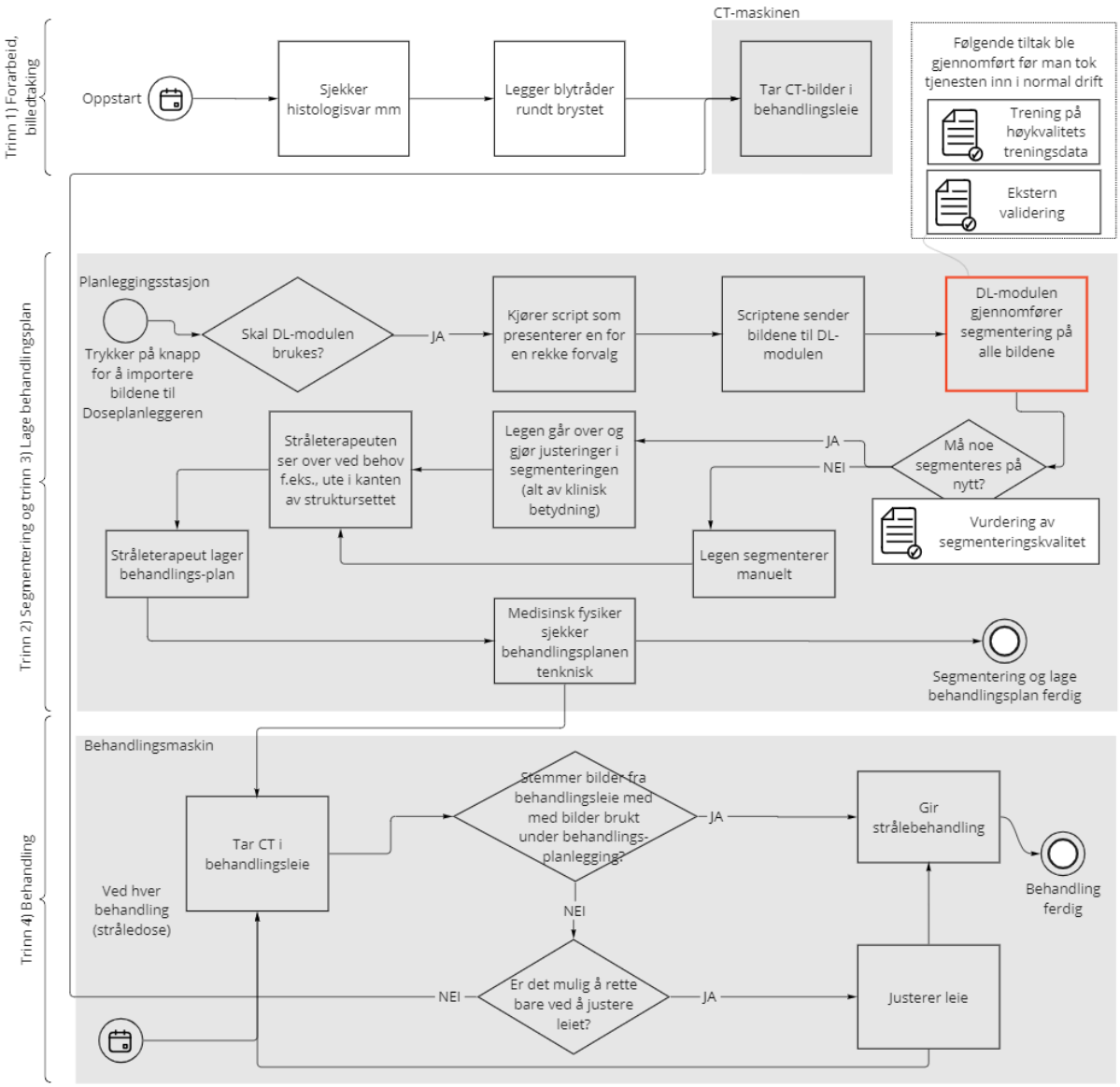
med blå strek rundt de rosa strekene. De tre store pilene + tilhørende tekst er lagt inn etterpå for å gi mer forklaring til leseren av studien.

Stråleterapeuten fortalte at han bruker å ta en kikk over etterpå: *"Modellen fungerer veldig bra, men den er ikke helt perfekt, slik at man kan se at enkelte strukturer på noen få bilder ikke er helt optimalt inntegnet, og da må vi inn og bruke manuelle tegneverktøy og justere pittelitt på strukturene før vi kan si oss helt fornøyde med struktursettet"*. Fysikeren fortalte at *"det kan være strukturer som legen ikke har så stort fokus på fordi at det kanskje ikke er så klinisk relevant".*

I det tredje steget genereres et forslag til behandlingsplan, som justeres og godkjennes av helsepersonellet. DL-modulen brukes ikke i den forbindelse, men gir gevinster i form av at *"man kommer jo raskere i gang med planleggingen"* (st).

Overordnet flytdiagram for den nye arbeidsprosessen

Figur 11 viser overordnet arbeidsflyt med bruk av DL-modulen. Trinnene har samme numre som i Figur 8. Trinn 1 er forarbeid og billedtaking. Merk at (1) helsepersonellet hadde, eller skaffet seg, mer kontekst for vurderingene som skulle gjøres, både gjennom prøvesvar, samtaler og at man la blytråd rundt brystet – noe som var til nytte for både legene og DL-modulen. (2) CT-bilder ble tatt i *behandlings*leie, som var viktig for å få høy kvalitet på behandlingen senere. Neste trinn, trinn 2, skjedde på planleggingsstasjonen, og her var det DL-modulen ble brukt. Her var det to nye beslutninger som måtte gjøres: Hvorvidt den DL-baserte modulen skulle brukes på en gitt pasientgruppe, og hvorvidt segmenteringen utført fra DL-modulen skulle beholdes som den var, rettes eller forkastes. Etter dette laget man en behandlingsplan basert på de segmenterte bildene, og gav strålebehandling (trinn 3 og 4).



Figur 11 Case 1: Oversikt over arbeidsflyt inkl. viktige kvalitetssikringstiltak

Figuren viser oversikt over arbeidsflyt inkludert viktige kvalitetssikringstiltak knyttet til bruk av DL-modulen. Grå bakgrunn viser hvilke trinn som gjøres på ulike maskiner/i ulike programmer. Bruk av DL-modulen er markert med litt tykkere, rød ramme (litt over midten av bildet, helt til høyre). Teksten på siden til venstre refererer til trinn-numre fra Figur 8. Merk at i enkelte tilfeller kan anatomien være vesentlig forandret, noe som medfører at man i nederste boksen (trinn 4) må starte på toppen med ny billedtaking.

4.1.4 Rammer for bruk

I bruk var det som allerede nevnt to nye hovedbeslutninger som måtte tas, knyttet til hvilket ansvar som skulle gis til den DL-baserte modulen i hvert enkelt tilfelle: Hvorvidt produktet skulle brukes på en pasient, og i hvilken grad segmenteringen som ble produsert av modulen skulle brukes, rettes eller erstattes. Her vil jeg beskrive ulike typer rammer for bruk som ble nevnt av intervjupersonene og som er relatert til disse to beslutningene. Jeg starter imidlertid med en kort beskrivelse av treningen av DL-modellen som er brukt, siden den har konsekvenser for rammene som jeg beskriver etterpå.

CT-bildene må være innenfor rammene av treningsdata

Modellen som ble brukt i DL-modulen ble ifølge stråleterapeuten trent på en *"meget presis inntegning på 200 pasienter"* noe som gjorde at *"den er veldig godt trent på det den gjør"* (st). Modellen var trent på vanlige, ganske standard pasienter, og for *"pasienter som er innenfor rammene av denne standardiseringen, så vil jeg si at denne modellen er – på sett og vis – bedre trent enn kanskje onkologen som da selvfølgelig må forholde seg til alt mellom himmel og jord av pasienter (…) Kanskje kan man si at modellen er bedre enn legen for pasienter som passer innenfor den rammen"* (st). I enkelte tilfeller kunne imidlertid CT-bildene være annerledes enn treningsdataene, og fysikeren fortalte at *"det helsepersonell har, er selvfølgelig en veldig god generell kunnskap som gjør at man takler – selvfølgelig – mye bedre situasjoner hvor man kommer utenfor boksen, utenfor rammene"* (f). Dersom man likevel bruktr DL-modulen på pasienter som var vesentlig forskjellige fra treningsdataene *"så kan man tenke seg at modellen vil bli litt forvirret og lage et feil resultat"* (f). Det som var utenfor det modellen kunne brukes til, kalte fysikeren *"utenfor rammene"*. Tabell 3 oppsummerer scenarier som var utfordrende fordi de var utenfor rammene, og hvordan man har gjort tilpasninger til dette.

Utfordrende scenarier	**Tilpasning**
Pasienter utenfor treningsdata (kvinner som er operert for brystkreft)	*"Vi konfigurerer i våre script hvilke pasientgrupper den skal kjøres på"* (f)

Pasienter med fjernet bryst eller implantat	Segmenteringsmodulen ble kjørt, men feilsegmenterte strukturer ble forkastet *"og tegnet manuelt". "Jeg forkaster noen av strukturene", men "ikke risikoorganene, og ikke lymfeknutestasjonene"* (o)
Manglende klips i sårhulen[20]	
Bolus	
Egenskaper ved bildene, som strukturer som ligger tett på hverandre og har lite kontrast	Segmenteringen kjøres, og feil korrigeres manuelt: *"Når vi er inne og tegner strukturene så er vi kanskje inne og gjør manuelle justeringer i window level for å klare å skille dem med det blotte øyet"* (st)
Finne nøyaktig grense for brystet på CT-bildene	Legger blytråder *"akkurat rundt der vi klinisk ser at brystet går" før billedtaking"* (l)
Avvikende liggestilling under billedtaking	*"Armen opp er standard"* (o)

Tabell 3 Case 1: Oppsummering av utfordringer og tilpasninger

Scenarier som er mer utførlig beskrevet i den videre teksten, er bare nevnt kort i tabellen.

Etter en brystoperasjon kan sårhulen fylles med væske ("serom"[21]), og stråleterapeuten fortalte at dersom pasienten får tappet den for væske, *"så har man anatomi som vil avvike fra det normale".* I enkelte tilfeller er brystet fjernet, og onkologen forklarte at da *"er ikke segmenteringen veldig god (…) Modellen er rett og slett ikke trent for det"*. I andre tilfeller er brystet rekonstruert, *"for eksempel dersom man har fått laget et silikonbryst".* Da *"vil silikonbrystet skille seg fra vanlig brystkjertelvev, så det også kan være en utfordring for modellen å gjenkjenne"* (st). Stråleterapeuten forklarte at i slike tilfeller *"vil jo vi [helsepersonellet] ha bedre evne, eller bedre datagrunnlag, for vi har sett dette mange ganger før, og kan da gjøre den manuelle justeringen som skal til for at det skal bli riktig"* (st).

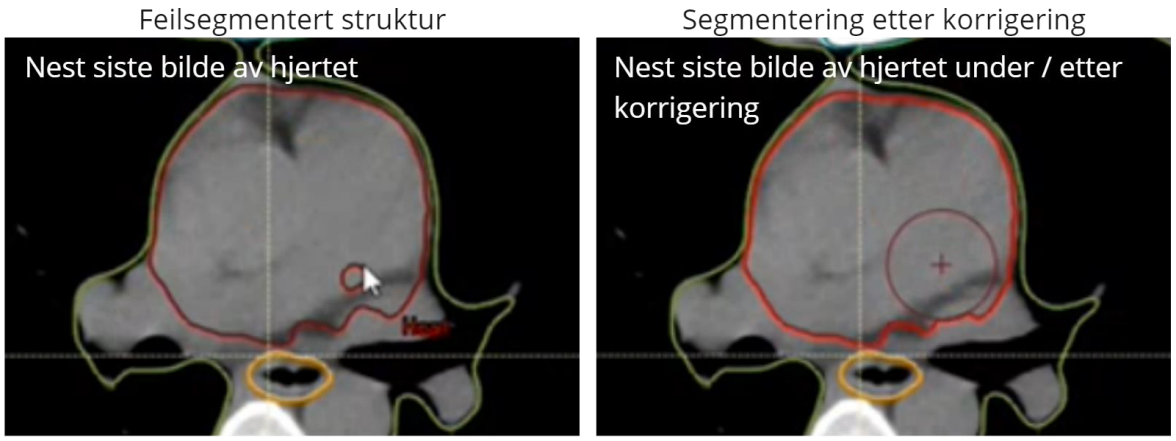
Når strukturer ligger tett og har lite kontrast kan **"***den ene strukturen blir definert litt over i nabostrukturen"*, fortalte stråleterapeuten. I tillegg er det enkelte snitt som ikke segmenteres riktig,

[20] Akkurat der kulen satt i brystet før den ble tatt ut, skal kirurgene i legge igjen små klips som viser godt igjen på bildene, men det er ikke alltid klipsene blir satt inn av kirurg

[21] Serom er en ansamling av sårvæske. Det kan for eksempel opptre i sårhulen etter operasjoner i bryst eller armhule.

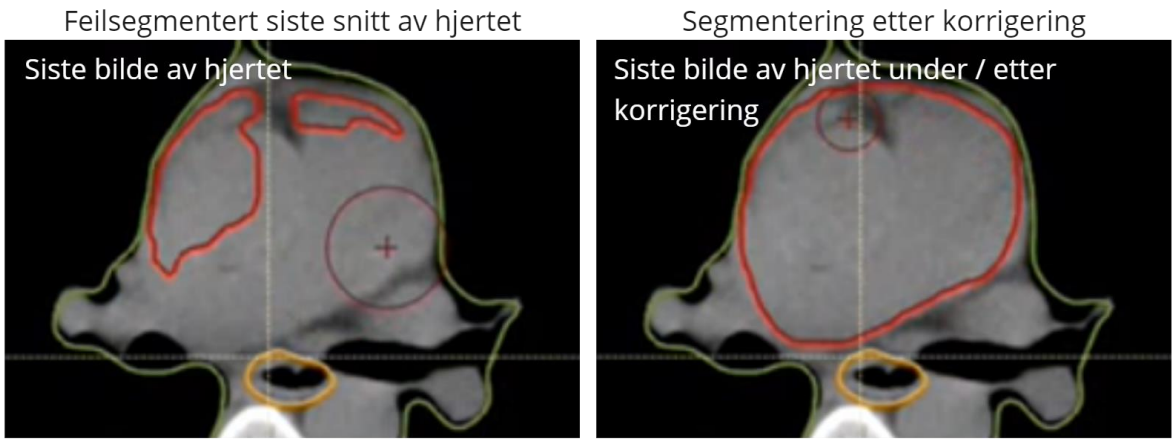
"det er typisk øverste og nederste snitt på hjerte, på brystet, og (…) det øverste lymfeknuteområdet", fortalte onkologen. Seksjonslederen fortalte at *"det er ett snitt der brystet stopper og glandulastasjonene fortsetter – der også er der en ting som stort sett alltid må korrigeres"* og *"nederst på esophagusen (…) kan du få en dobbeltstruktur".*

Figur 12 - Figur 14 viser eksempler på feilsegmenteringer. Bildene til venstre er før korrigering, og bildene til høyre er etter. Strukturene er markert med ulike fargekoder slik at helsepersonellet lett skal kjenne dem igjen.



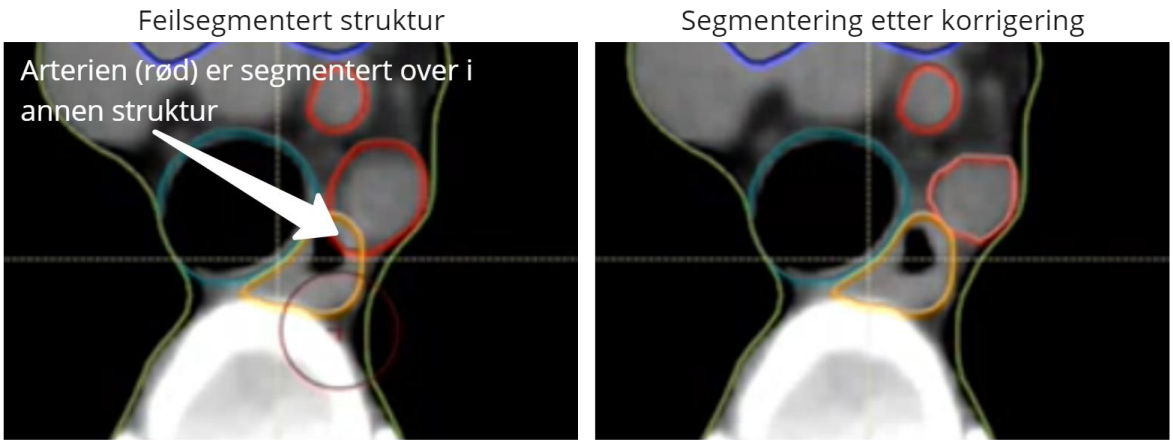
Figur 12 Case 1: Eksempel på feilsegmentering av hjertet (1)

Eksempel 1 (øverst til venstre) viser at hjertet (rødt) har fått en ekstra segmentering inni seg (rød, liten sirkel, markert med pil/musepeker). I bildet til øverst til høyre bruker helsepersonellet korrigeringsverktøy (rød sirkel med kryss i) for å ta bort denne ekstrastrukturen.



Figur 13 Case 1: Eksempel på feilsegmentering av hjertet (2)

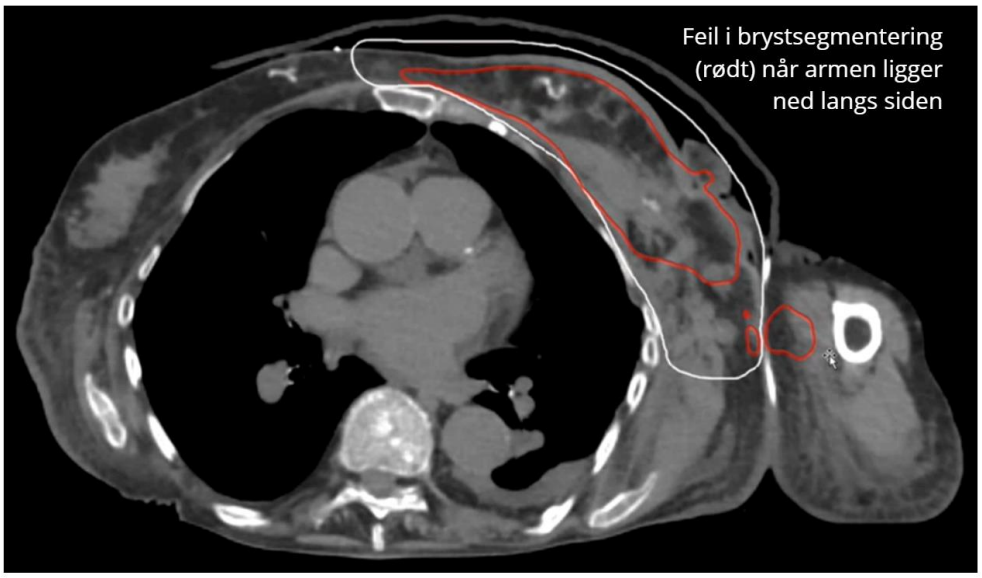
Eksempel 2 viser siste bildet/snittet av hjertet. I stedet for å vise dette som en struktur, er den feilaktig delt opp i to mindre strukturer (i rødt). Rød sirkel med kryss i er korrigeringsverktøy. Bildet til høyre viser at man bruker korrigeringsverktøyet for å samle de to strukturene til en og i tillegg ta med den delen av hjertet som ikke ble med.



Figur 14 Case 1: Eksempler på feilsegmentering av strukturer som går over hverandre

Eksempel 3 viser at spiserøret (i gult) og en arterie (i rødt) overlapper. På bildet til høyre er dette korrigert slik at arterien ikke overlapper med spiserøret.

Ifølge fysikeren er pasientens stilling under billedtakingen viktig for korrekt segmentering. *"Alle pasientene i dette datamaterialet har for eksempel armene sine opp, men så har vi sett at dersom vi har en pasient som har armen ned istedenfor, og vi prøver å kjøre modellen likevel, så blir modellen forvirret og lager et litt dårlig resultat"*. Onkologen forklarte at dette skyldtes at *"anatomien i lymfeknuteområdet endrer seg når armen har en annen posisjon".*



Figur 15 Case 1: Feil i brystsegmentering når armen ligger ned langs siden

Vi ser at det som er markert som bryst her blir litt "rotete", med strukturer markert ute i armen (små røde sirkler til høyre i bildet).

I sjeldne tilfeller brukes en bolus under billedtakingen for å få en ekstra stråledose ut i huden. Stråleterapeuten forklarte at da *"legger man en pute på pasientens hud, som er en halv cm i tykkelse som er av et materiale som er vevsekvivalent, for at man skal få en doseoppbygging som gjør at man får full stråledose helt ut i hudoverflaten. Og da blir det en struktur også som man ikke har trent modellen på og som kan forvirre modellen litt"* (st).

Får hjelp av markører

For å hjelpe både helsepersonellet og DL-modulen brukes markører som vises igjen på CT-bildene. At helsepersonellet legger blytråder rundt det aktuelle brystet på pasienten er ett eksempel som allerede er nevnt. Et annet eksempel på markør er at det legges klips i sårhulen: *"Når en skal få autodefinert det som vi kaller tumorseng, eller akkurat der kulen satt i brystet før den ble tatt ut, så skal kirurgene i utgangspunktet legge igjen noen klips i sårhulen under operasjon, altså noen små klips som viser veldig godt igjen på bildene våre. De er til stor hjelp for modellen når den skal klare å definere eget volum med omriss rundt der kulen satt".* Det er noe variasjon i praksis ved operasjon, som får konsekvenser ved segmentering, for *"det er ikke alltid klipsene blir satt inn av kirurg"*, og *"modellen strever litt mer når den ikke har de klipsene"* (st).

Brukes kun av erfarne

Onkologen fortalte at også erfaring har mye å si for om de får lov å bruke segmenteringen fra produktet: *"Vi har jo assistentleger hos oss, eller leger i spesialisering, og de skal jo trenes i å tegne målvolum, så de bruker ikke modellen for de er nødt til å lære seg det manuelt (…) De har ikke nok erfaring med hva som er fasiten, så de kan ikke bruke produktet etter min mening (…) Vi [erfarne] bruker produktet, men jeg kontrollerer jo produktet på en måte".*

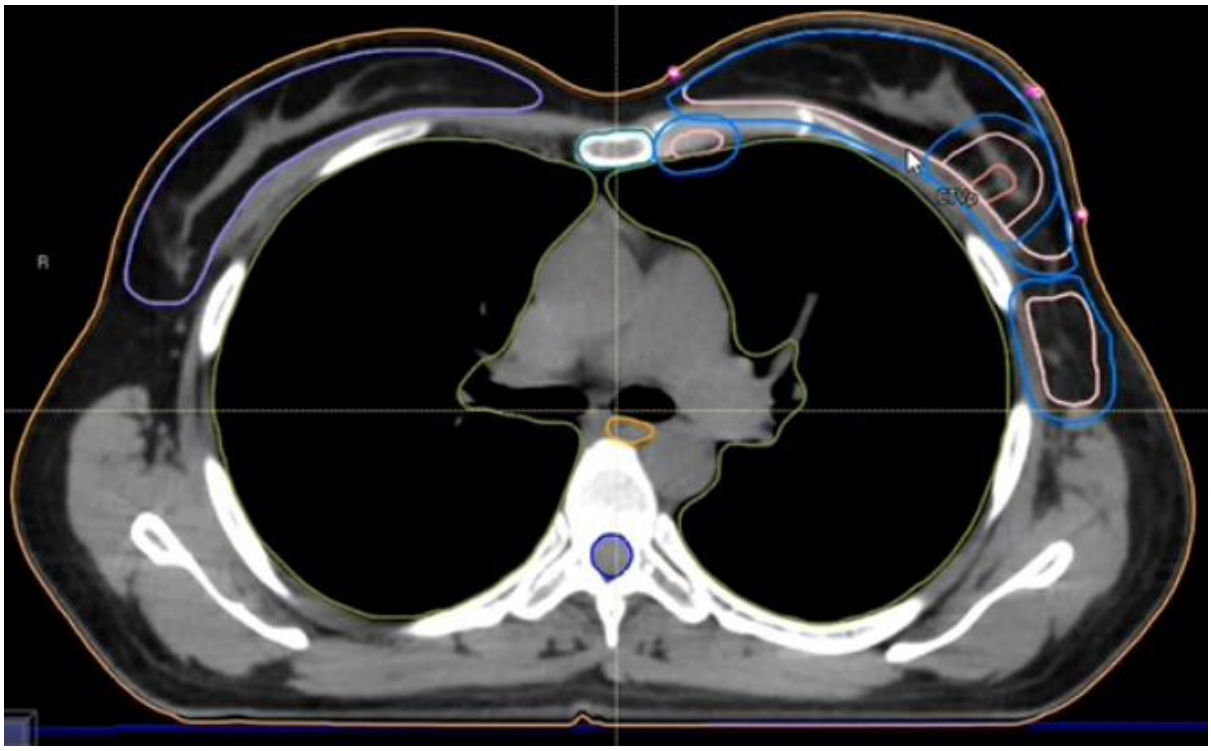
Innenfor rammene beskrevet i avsnittene over stolte helsepersonellet til en viss grad på resultatet fra modulen. Onkologen fortalte at *"jeg stoler på den med forbehold",* mens fysikeren mente at *"kanskje kan man si at modellen er bedre enn legen, for pasienter som passer innenfor den rammen".* Selv om arbeidet var innenfor rammene av hva de mente modulen kunne håndtere, så beskrev fysikeren resultatet fra modulen som *"et forslag"* til segmentering, og seksjonslegen fortalte at *"vi kontrollerer det alltid"*, noe fysikeren bekreftet: *"Til syvende og sist vil legene gå gjennom og kvalitetssikre resultatet, og gjøre de justeringene som trengs (...) Vi vil jo aldri bruke denne teknologien blindt - å lage en behandling uten at noen mennesker har vært innom og godkjent og justert resultatet".*

4.1.5 Erfaringer fra innføringen

Rask og bra kvalitet

Enhet 1 hadde brukt produktet i underkant av et halvt år da intervjuene ble gjennomført, og erfaringene er derfor basert på en relativt kort bruksperiode. Alle intervjupersonene var svært positive til DL-modulen og fortalte at den innebærer store besparelser for den pasientgruppen det brukes på. Stråleterapeuten sammenfattet det slik: *"Jeg vil si ja, med to streker under, det er blitt kjempebra"*. Et hovedmål for innføringen var å spare tid – uten å redusere kvaliteten, og ifølge seksjonslederen er DL-modulen både rask og nøyaktig: *"Det som overrasket oss, er at det er så god kvalitet på segmenteringen etter så ekstremt kort tid (…) Det er enorme besparelser"* (sl). DL-modulen gjør ferdig segmenteringen av en pasient på *"bare på ca ett minutt, og så vil da legene trenge å gå over og kvalitetssikre resultatet, så totalt bruker de kanskje opp til 10 minutter"* (st). *"I utgangspunktet vil det spesielt være tidsbesparende for legeteamet vårt, siden det er de som tegner inn volumet som skal ha strålebehandling, og det betyr at legen i stedet for å sitte bak en PC og tegne bilde for bilde, får tid med pasienten"* (st), men *"det er også veldig tidsbesparende for oss som skal sitte og doseplanlegge, vi slipper også å bruke tid på å tegne inn alt av risikoorganer, som er de*

strukturene vi ønsker å skåne for strålebehandling" (st). *"Og så har modellen noen åpenbare fordeler. Litt flåsete sagt, men modellen blir aldri trøtt og sliten, opplever aldri samtidskonflikter som å svare på calling og telefoner, og har heller aldri fulle poliklinikklister. Den gjør alltid det den er trent opp til å gjøre – hverken mer eller mindre"* (sl).



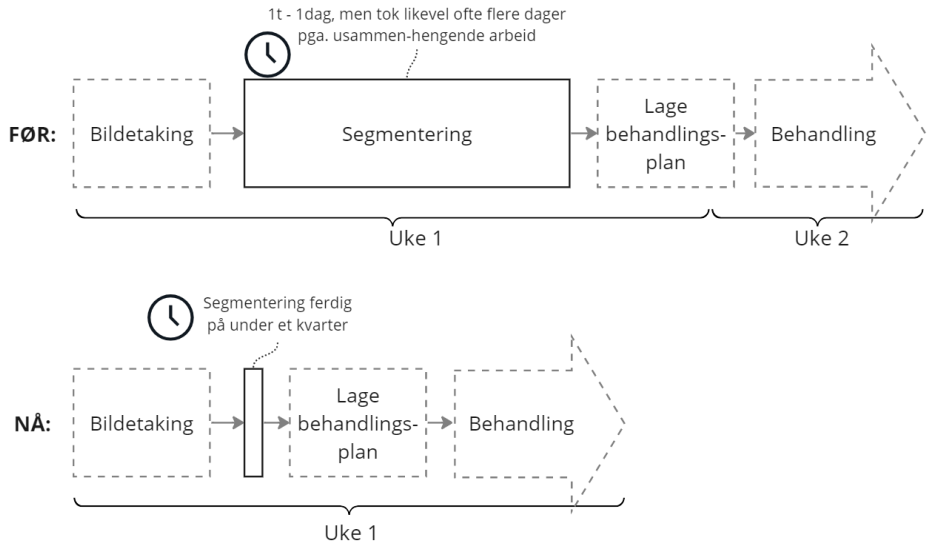
Figur 16 Case 1: Eksempel på segmentert snitt med DL-modulen

Målvolumene er markert med blå streker (rundt rosa streker) oppe til høyre i bildet (der det også er en liten hvit pil/musepeker)

Sparer pasienter for venting og reiser

Raskere segmentering sparer pasientene for venting og reiser. Figur 17 viser et eksempel: Pasienter som skal ha behandling i 5 dager måtte tidligere inn og ta CT en uke før selve behandlingsoppholdet, men siden segmenteringen nå gjøres så raskt, så kan man vente med CT til de kommer inn for behandling: *"La oss si at de kommer inn til avdelingen kl 10, så tar vi en CT. De får en legekonsultasjon klokken 11, og mens de har en legekonsultasjon, så har vi jo allerede startet behandlingsplanlegging på grunn av at modellen basert på kunstig intelligens har gitt oss alle*

volumene vi trenger" (st). Da kan de få første strålebehandling mandag og siste behandling fredag før de reiser hjem.



Figur 17 Case 1: Skjematisk fremstilling av tidsbruk før og nå

Den totale ledetiden har gått ned. Merk at størrelsen på boksene er symbolsk, og viser ikke korrekt bilde av verken tidsbruk på hvert prosesstrinn, eller hvor mye tidsbruken er redusert i dag.

Trygge på bruken

Helsepersonellet har samarbeidet med produsenten og segmentert store deler av treningsdataene selv. Dette så ut til å gjøre dem positive til produktet og trygge på rammene for bruk av produktet. Seksjonslederen fortalte at *"den [DL-modulen] gjør bare det den er programmert til å gjøre. Vi har veldig styring, og det er jo veldig behagelig. Vi vet (…)"* Når de fortalte om styrker og svakheter ved produktet, forklarte de det gjerne ved hjelp av detaljert kjennskap til treningsdataene. Et eksempel er da onkologen fortalte om en pasient som hadde fjernet brystet: *"Den enkle og gode forklaringen på hvorfor segmenteringen er dårlig på de, er at modellen er rett og slett ikke trent for det, for det var veldig få pasienter som var rekonstruert eller hadde ablatio[22] i materialet[23] vårt"* (o).

[22] Ablatio = fjerning av en kroppsdel, her brystet
[23] Data*materiale*, det vil si treningsdataene

Intervjupersonene hadde tillit til produktet, men presiserte at de ikke kunne stole hundre prosent på resultatet. Seksjonslederen forklarte at *"de fleste prosjekter møter en viss motstand i starten. Det var også tilfelle i dette prosjektet. Vi var alle lett skeptiske til dette, men ble raskt overbevist når vi så hvor godt modellen faktisk fungerte".* Videre fortalte han at *"med årene kommer vi til å stole enda mer på at outputen er hundre prosent"*. Onkologen fortalte at *"litt finjustering av strukturene er det likevel behov for, og det er uansett en kvalitetssikring vi er nødt til å gjøre før vi iverksetter en dosebehandling, vi kan ikke stole blindt på det, i hvert fall ikke nå i starten. Vi har jo ikke brukt det så lenge".* Onkologen fortalte at de nå har fått erfaring og vet hvilke strukturer som trolig må rettes: *"Da vet jeg jo hvor det som oftest er litt feil, hvis man kan kalle det det, der hvor modellen ikke er helt optimal"*. Ikke alle bruker DL-modulen, og stråleterapeuten problematiserte et dilemma knyttet til kompetanse: *"Hvordan skal man som utrent vite hva som er godt nok – det er et dilemma når man innfører sånne modeller for de som ikke har det helt i ryggmargen – hvordan slikt skal se ut? (…) Det er en problemstilling som går på opplæring"* (st).

Små endringer

De ansatte opplever at det er fint å være i front på fagområdet: *"Det er litt kult å ha en modell som kan selges over hele verden og så kan vi kalle den litt <Sykehusnavn>sk"* (sl). Utviklingen og innføringen medførte en rekke oppgaver eller endringer, fra prosjektrelaterte oppgaver knyttet til å lage treningsdata og kvalitetssikre produktet, teknisk integrasjon og prosessmessige endringer. Til tross for dette er de ansattes opplevelse at det var lite endringer, og at arbeidsflyten var omtrent som før. Selv om produktet har redusert tiden brukt på segmentering av brystkreftpasienter, var alle intervjupersonene også enige om at bruk av DL-modulen ikke har endret arbeidshverdagen deres i særlig grad, siden brystkreftpasienter utgjør en liten del pasientene ved enheten. Fysikeren fortalte at han årlig hadde *"i hvert fall et par hundre [pasienter]. Men det er pasienter av alle slags typer, ikke sant, det er prostata, hode, lunge, bryst og alt mulig".* Stråleterapeuten sa det samme: *"Arbeidsdagen min er ikke spesielt endret. Jeg har de samme arbeidsoppgavene, men jeg kan arbeide mer effektivt for akkurat denne her pasientgruppen"*.

Planlegger utvidelse av bruk av KI

Erfaringene fra utviklings- og innføringsarbeidet skal brukes til å utvide bruken eller utvikle modeller for andre kreftpasienter. Ett av målene er å *"gjøre en såkalt adaptiv, eller skreddersydd, strålebehandling i fremtiden"* (f). Det vil si at man tar nye CT-bilder hver gang pasienten skal ha

behandling, slik at bildene passer helt med *"dagens anatomi"*[24] hver gang. *"Da kan man i prinsippet gi strålebehandling som enten har mindre bivirkninger eller høyere effekt enn i dag"*, men *"i dag kan det være litt utfordrende hvis man trenger å bruke veldig lang tid på å tegne inn slike strukturer på nytt"*. *"Det vi drømmer litt om i en liten forlengelse av dette, er at vi ønsker å tilby skreddersøm strålebehandling hver dag, til pasienter, også innenfor de samme ressursene som vi besitter"* (sl). Personellet ønsker å utvide teknologien til andre diagnosegrupper. *"I utgangspunktet er det pasienter med brystkreft og regionale glandula, men du kan egentlig bruke modellen på mange andre strukturer i thorax-området", og "vi er godt i gang med nye modeller allerede, i andre regioner*[25]*"* (sl). De har startet et tilsvarende arbeid for prostatakreft. Prostataen ligger mellom blæren og endetarmen som ofte endres mye i størrelse på kort tid, og dytter på omliggende strukturer, noe som endrer hvor strålebehandlingen burde rettes inn. Kvalitetsgevinstene er derfor forventet å bli større for prostatakreft.

4.1.6 Oppsummering

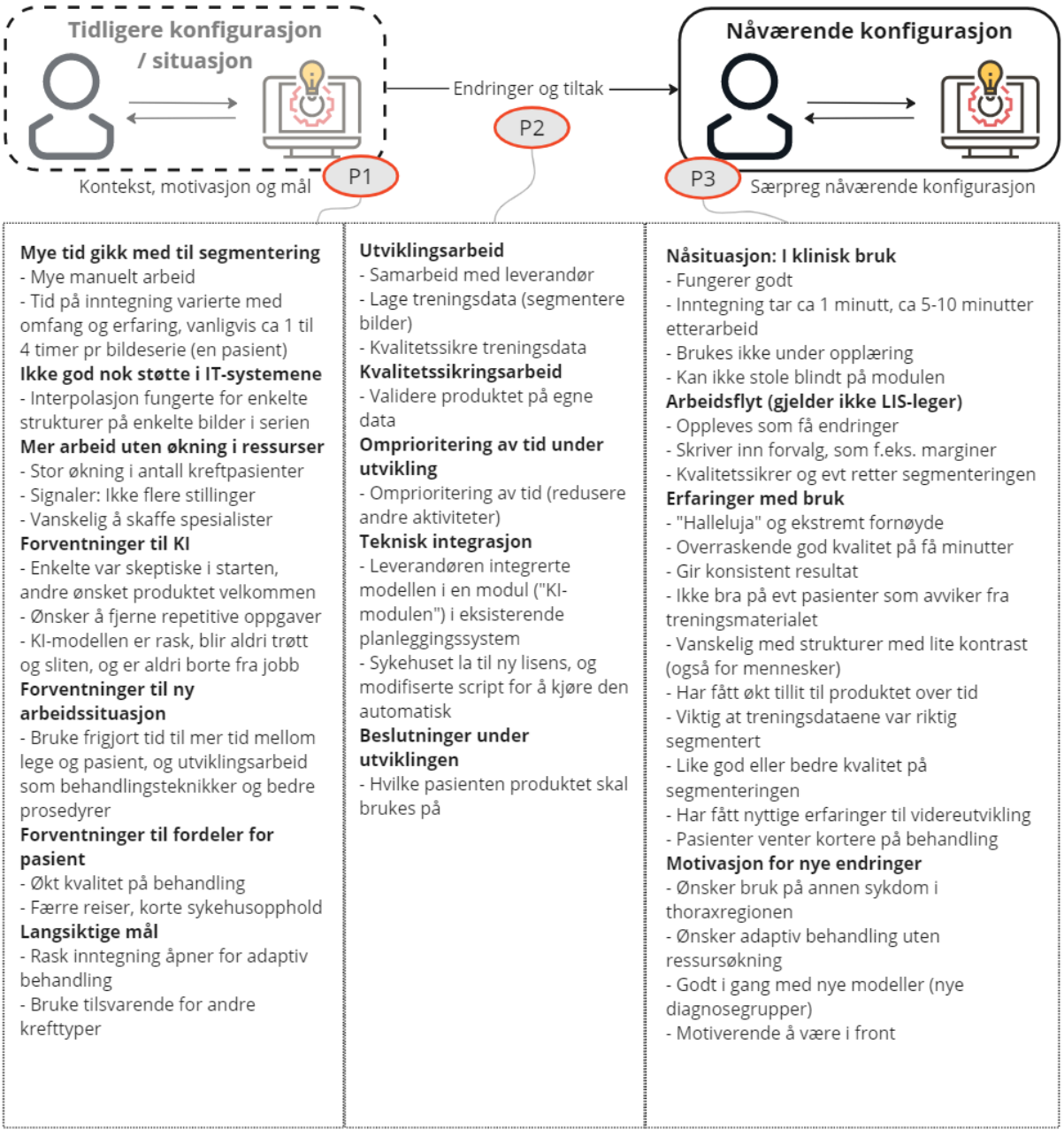
Enhet 1 bruker en DL-modul for segmentering av CT-bilder ved brystkreft. Arbeidet ble startet opp som en reaksjon på at de har stor økning i antall kreftpasienter, mye tid gikk med til manuell segmentering og eksisterende IT-systemer gav ikke god nok støtte. Produktet er trent på et relativt begrenset treningsmateriale med standard pasienter, men med høy kvalitet på segmenteringen. De har samarbeidet tett med leverandøren og et annet norsk sykehus, og har i den forbindelse laget halvparten av de segmenterte treningsdataene samt kjørt validering etter trening. Det tekniske implementeringsarbeidet har vært av begrenset omfang, siden leverandøren har integrert produktet som en modul i programvare som de allerede bruker. I bruk har produktet medført to nye beslutninger: Hvilke pasienter skal produktet brukes på – dette har de lagt inn i et script slik at de slipper å vurdere det i hvert tilfelle, og hvorvidt resultatet fra produktet skal brukes som det er, rettes eller ses bort fra. De er svært fornøyde med produktet. Segmenteringen er langt mer nøyaktig enn tidligere teknologisk støtte for dette, og tiden har gått ned fra minimum 1t pr pasient til maksimum 10 minutter pr pasient, inkludert menneskelig kvalitetssikring. DL-modulen har enkelte

[24] Her menes det at man kan tilpasse behandlingen til hvordan organene ligger ved hver behandling. Plassering vil være litt forskjellig fra behandling til behandling. Forskjellene vil være større ved enkelte andre kreftformer, f.eks. prostatakreft, som ligger nær urinblæren

[25] Andre regioner av kroppen

svakheter når pasientene ikke ligner treningsdataene, noe som gjør at all segmenteringen kvalitetssikres og enkelte snitt må rettes. Opplevelsene av den nye modulen er positiv og de har derfor satt seg nye mål, både for denne pasientgruppen og for nye pasientgrupper.

Figur 18 oppsummerer faktorer som ble nevnt for hver av delproblemstillingene, også utover de som jeg har skrevet mer utfyllende om i teksten over.



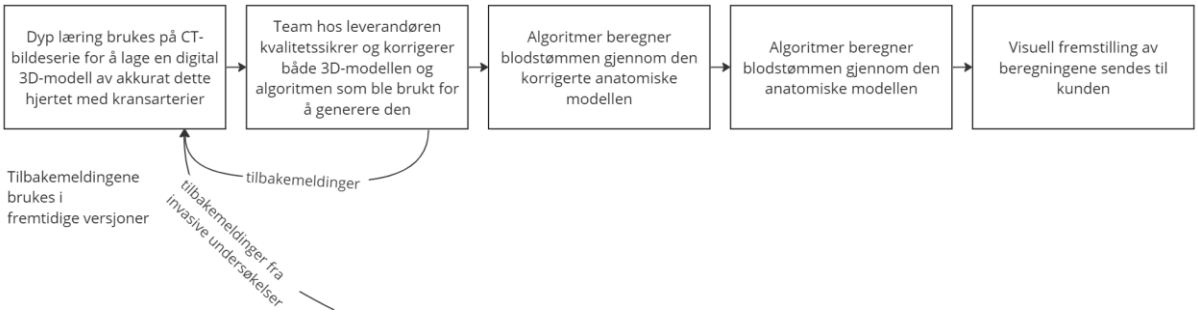
Figur 18 Case 1: Oppsummering av delproblemstilling 1-3

4.2 Case 2 DL-basert analyse av angiogrammer

Enhet 2 er en enhet for angiografi som har brukt en tjeneste basert på dyp læring (heretter kalt "DL-tjenesten") siden høsten 2017. Da intervjuene ble gjennomført var tjenesten blitt benyttet på et par hundre pasienter. Tjenesten brukes på pasienter som har brystsmerter og legen har derfor mistanke om sykdom[26] i kransarteriene, blodårene som forsyner hjertemuskelen med blod. Pasientene blir derfor sendt til sykehuset for å ta koronar angiografi, en CT-basert undersøkelse for å diagnostisere denne typen sykdom. DL-tjenesten brukes bare på ikke-akutte pasienter som røntgenlegen ut fra CT-bildene *tror* er friske, men hvor han trenger å være helt sikker. Røntgenlegen (rl) og intervensjonskardiolog (ik) er primære brukere av DL-tjenesten. I tillegg har jeg intervjuet IT-radiograf (ir) og radiograf (r).

DL-tjenesten som sykehuset bruker eies av et selskap i USA og brukes av mer enn 400 sykehus i verden. Figur 19 viser hovedtrinnene i prosessen hos leverandøren av tjenesten. Det bruker proprietære dyp-læring-algoritmer til å lage en digital 3D-modell av kransårenes anatomi ut fra tilsendte CT-bildeserier. Hver 3D-modell som genereres av DL-systemet kvalitetssikres og korrigeres av et tverrfaglig analyseteam, bestående av personer med teknisk eller medisinsk bakgrunn. Rettingene deres brukes både umiddelbart for å få en så korrekt 3D-modell for den konkrete pasienten som mulig, og som treningsdata for fremtidige versjoner av programvaren. Når 3D-modellen er rettet slik at den stemmer best mulig overens med CT-bildene, brukes pasientspesifikk *computational fluid dynamics (CFD)* for å beregne blodstrømmen gjennom den anatomiske modellen. Visuell fremstilling av blodstrøm gjennom kransarteriene blir til slutt sendt tilbake til kunden (sykehuset).

[26] Dette kan typisk være stenose, altså fortetninger i blodårer



Figur 19 Case 2: Hovedtrinn i kvalitetssikringsprosessen hos leverandøren

4.2.1 Konteksten for ønske om endringer

Røntgenlegen fortalte at *"akkurat her, hvor jeg jobber, på røntgenavdelingen her, så sitter jeg ganske alene om feltet jeg holder på med, som er hjerte-CT. Det er ikke masse kolleger å ta av". "Hadde jeg hatt et stort miljø, så er det ikke sikkert jeg hadde hatt like mye behov for det [DL-systemet]"*. Røntgenlegens arbeid er å studere CT-bildene av hjertet og vurdere hvorvidt det er tegn til sykdom. Dersom han ikke ser tegn til sykdom, så er det stor sannsynlighet for at hjertet har god blodforsyning. Ser det derimot ut som at det kan være sykdom, *"så er det bare rundt halvparten, eller litt under det, som har sykdom som trenger behandling (…) og motivasjonen her var da å prøve å få ned den andelen pasienter som kom til [unødig] koronar angiografi etter CT-undersøkelse" (ik). Det er flere fordeler med å sørge for at pasienter ikke kommer til unødige undersøkelser, som "å redusere plager og risiko for pasienter"* (ik). Når analysen av CT-bildene er ferdig, *"da har pasientene som oftest forlatt sykehuset"* (rl), og en avgjørelse om koronar angiografi betyr derfor en ny sykehustur en annen dag, noe som før påførte pasienten ulemper og helseforetaket ekstra kostnader. *"Hos oss er det gjerne sånn at når pasienten skal til undersøkelse her, så er det en kostbar reise, og gjerne også i forbindelse med opphold på hotell, så de beløpene blir ofte større enn det produktet koster (…) Det vil jeg tro var hovedgrunnen til at vi fikk lov til å prøve ut det her"* (rl).

Det er spesielt to situasjoner hvor det var vanskelig å være sikre på egne vurderinger: *"CT-undersøkelsen kan være vanskelig å tolke hvis det er kalk i blodåreveggen, fordi kalk har samme farge som kontrast[27], og kontrast inni årene er det vi bruker når vi vurderer størrelsen på karene og*

[27] *Kontrast* brukes her i betydningen "kontrastvæske", som man sprøyter inn i årene for lettere å se forskjell på blodårene og blodet på bildene

blodårene" (ik) og *"det andre er dersom det ikke nødvendigvis er trangt, men lengre område der det er sykdom i årene der vi er usikre på om den lettgradige trangheten totalt sett kan gi en reduksjon i blodstrømmen"* (ik). Siden bare halvparten faktisk er syke, så var motivasjonen *"å få ned den andelen pasienter som kom til koronar angiografi etter CT-undersøkelse – å kunne avklare de pasientene uten at de trengte en invasiv[28] undersøkelse der man går inn i kroppen (…) Det er jo for pasientens beste om man kan avklare disse problemstillingene uten å måtte gå inn i kroppen, og det er jo nettopp det vi gjør med CT og <DL-tjenesten>"* (ik). Selv om invasiv koronar angiografi er en trygg prosedyre, så medfører det en liten risiko for *"blødninger (…), hjerneslag, risiko for hjerteinfarkt eller andre katastrofale hendelsen i karsystemet som man er inne i"* (ik).

4.2.2 Preimplementering og implementering

Preimplementeringsfasen og implementeringsfasen i case 2 gikk ut på å lese studier om produktet, sikre at CT-bildene var i henhold til leverandørens krav, og gjennomføre en innføringsperiode på noen måneder hvor man sammenlignet egne resultater fra invasiv undersøkelse med resultater fra DL-tjenesten. Her gjennomgår jeg dette arbeidet.

Studier, krav og innføringsperiode

Intervensjonskardiolog fortalte om hvordan de brukte studier i forbindelse med beslutning om å ta tjenesten i bruk. *"De få studiene som var gjort var ganske klare på at dette var et godt produkt"*. De ansatte følte likevel på en viss usikkerhet, og satte derfor opp en studie som de gjennomførte i innkjøringsfasen. I denne perioden ble resultatene fra både røntgenlegen og DL-tjenesten sammenlignet med resultater fra invasiv undersøkelse, *"så i starten gjorde vi begge deler[29] for å se om vi syntes dette var et godt produkt og om det var trygt"*, *"og de første månedene viste at det var veldig god overensstemmelse mellom resultatene fra kunstig intelligens og (…) invasive undersøkelser, og etter den innføringsfasen så gikk vi over til en produksjonsfase der vi stolte på de resultatene som kom fra den"*.

[28] *Invasiv* betyr i denne konteksten at man fører instrumenter inn i blodårene for å undersøke dem innenfra
[29] Begge deler betyr her at de sendte bildene til den DL-baserte tjenesten og gjennomførte i tillegg koronar angiografi

Røntgenlegen fortalte at tjenesteleverandøren har satt *"noen kriterier på hva som må være oppfylt for at vi skal kunne sende bilene til dem". "Det er at kontrasten i karene må ha en viss intensitet, den kan ikke være for svak, for utvannet, det må være en viss intensitet. Det skal ikke være for mange stenter[30] i karene, for eksempel, men den gruppa har vi ikke, uansett. Det skal ikke være kalk i karene som er som en ring". "Og før vi tok det i bruk, så måtte vi sende 10 undersøkelser til dem, sånn at de kunne se på vår kvalitet eller om vi måtte gjøre endringer for at de kunne godkjenne oss, for å si det sånn".*

Røntgenlegen fortalte at *"de [tjenesteleverandør] skal avvise hvis bildekvaliteten ikke er god nok, så de lager ikke noe tull på dårlige bilder". "Er kvaliteten redusert, så strever jeg i tolkningen, men da ville nok også <navn på DL-tjeneste> strevd med bildene, men muligens mindre enn jeg ville strevd med dem, for jeg bruker en annen teknologi – jeg bruker øynene". "Bildekvaliteten er ikke alltid god på CT-bilder. Det er helt avhengig av pasienten, og noen rører på seg, noen puster, og pulsen – uregelmessig puls, tykkelse på pasienten. Alt det her påvirker bildekvaliteten og da er det vanskeligere å tolke"*. Røntgenlegen fortalte videre at de hadde tett kontakt med leverandøren i starten: *"Tidligere var det representanter som kom reisende opp annenhver måned, og vi gikk gjennom undersøkelser som spriket, men de har sluttet å komme. Det har vel hatt noe med økonomi å gjøre også"*.

Intervensjonskardiolog fortalte at det i oppstartfasen ble gjennomført undersøkelse av i hvilken grad DL-tjenesten avvek fra resultatet etter invasiv undersøkelse, og at *"det var veldig få"* tilfeller. Imidlertid er det *"noen tilfeller der man ikke får avklart med dette produktet – der vi får svar at dette går ikke an å avklare, eller at man får ikke analysert. Og da ønsker de gjerne tilbakemelding på endelig resultat[31]"*. I tillegg ønsker de tilbakemeldinger, fortalte han, *"i de tilfellene hvor man skulle oppdage at det er diskrepans mellom svaret man får fra kunstig intelligens-produktet, og det som vi eventuelt finner på en avklaringsundersøkelse her" (ik)*.

Teknisk integrasjon og leverandørkontakt

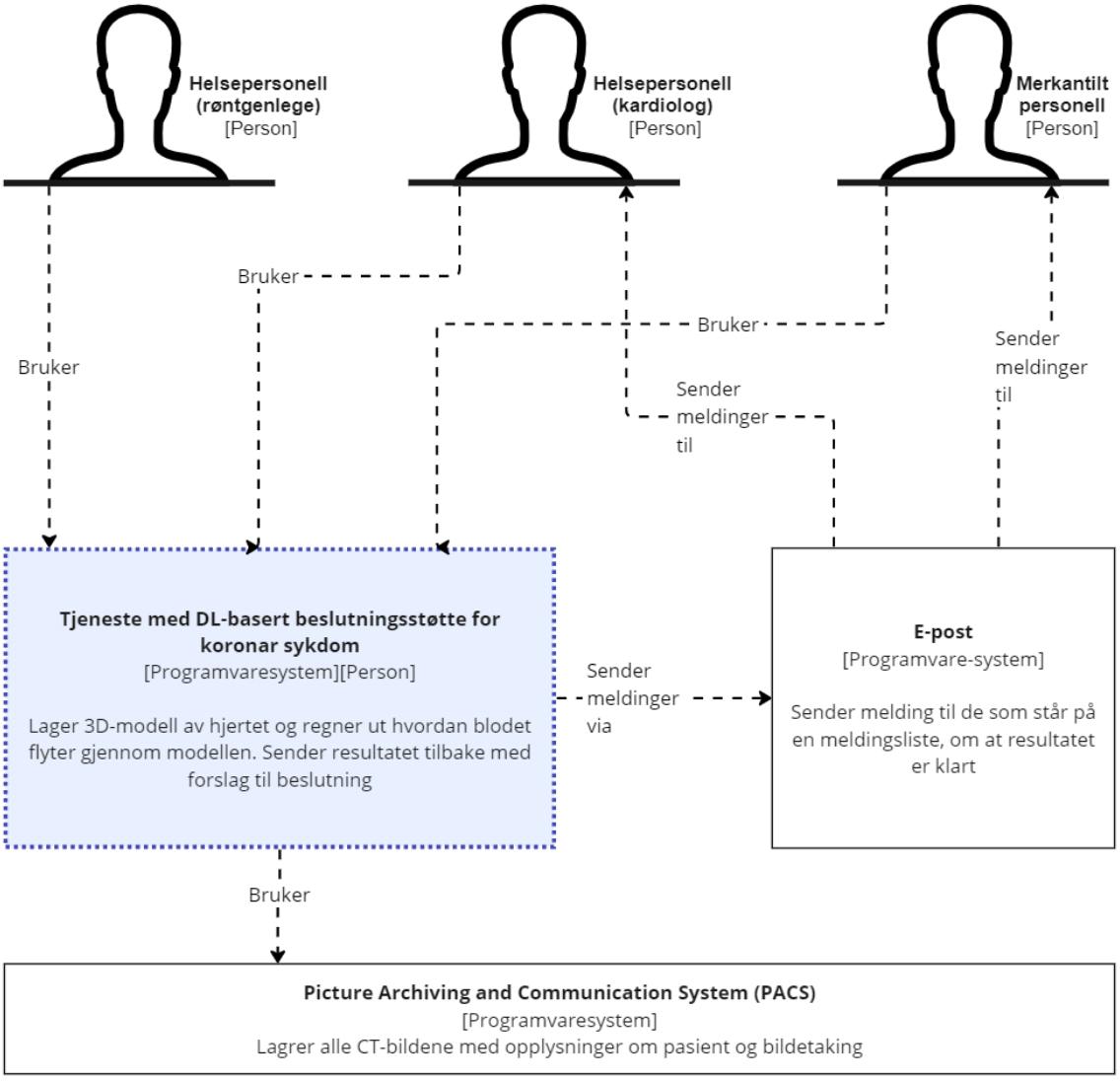
Når det gjelder tekniske endringer, har jeg ikke gått i detalj, men sett på *hvilken type* endringer som ble gjort.

[30] Stent er et lite rør som legges inn i blodårene for å holde det åpent
[31] Med "endelig resultat" mener han resultatet etter invasiv undersøkelse

Intervensjonskardiolog fortalte at første prosesseringen av CT-bildene skjer i et annet produkt, og *"det er der man gjør den første vurderingen etter CT-undersøkelsen".* Røntgenlegen utdypet at *"der kan man snurre rundt på hjertet, og dele det opp"*, og *"hvis man der ikke får avklart [tilstanden], så tar man ut fila i råformat og sender til <navn på DL-tjeneste>"* (ik). Røntgenlegen fortalte at de i røntgensystemet har funksjonalitet for å sende CT-bildene til ulike steder *"det er en lang liste, og der står det <navn på DL-tjeneste> så da klikker jeg på den og så klikker jeg på bildene jeg vil sende over (…) Det er noen få tastetrykk og så er undersøkelsen sendt over"* (rl). I produktet helsepersonellet bruker for prosessering av CT-bildene og 3D-visning av hjertet var der allerede på plass *"en egen protokoll på maskinen som heter <tjenestenavn>"* (r). Teknisk integrasjon var derfor ikke nødvendig. Det var bare behov for lisensnøkkel for denne tjenesten. Radiografen fortalte om integrasjonen at *"det er bare en standard node, som vi kaller det. Skal du sende det til Rikshospitalet, eller skal du sende det til <tjenestenavn>".*

Røntgenlegen fortalte at man går til "en nettside som man kan logge på med personlig bruker, og der får man resultatene". Det er få som har tilgang til dette: *"Resultatet ligger på en nettside som vi foreløpig bare er to her – det er jeg og han som har tilgang til den nettsiden – der vi kan logge inn, og så får vi det resultatet presentert der som en 3D-modell som kan roteres på"* (ik).

Figur 20 visualiserer konteksten som DL-tjenesten fungerer i. Tjenesten selv er plassert i midten og markert med lys blå bakgrunn. Over tjenesten er ulike brukere plassert. Under er et system som tjenesten bruker for å få tilsendt bildene. Til høyre er et system som tjenesten bruker for å distribuere svar.



Figur 20 Case 2: Systemkonteksten for DL-tjenesten

Pilene viser prosessflyt, og skal leses i pilens retning, som for eksempel: "Helsepersonell (røntgenlege)" + "bruker" + "tjeneste med (…)" . Merk at jeg i denne figuren også har tatt med merkantilt personell for å få helhet i figuren, men den sosiotekniske (re)konfigurasjonen som jeg studerer er knyttet til legenes (røntgenlege og invasiv kardiolog) bruk av systemet, siden de bruker tjenesten relativt likt.

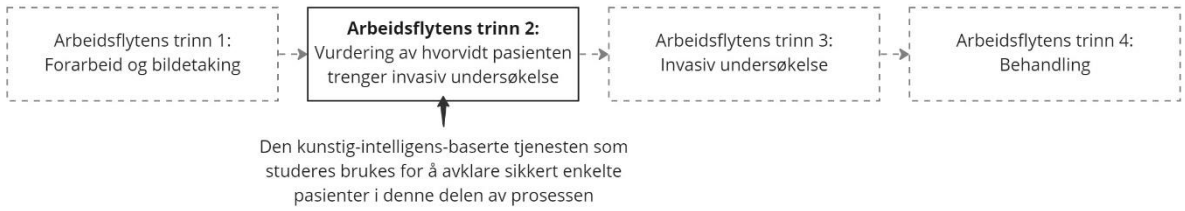
Røntgenlegen fortalte at DL-tjenesten, akkurat som han selv, ikke har noen pasientopplysninger, og bruker kun CT-bildene som datagrunnlag. Imidlertid er det et kvalitetssikringstrinn hos produsenten hvor *"teknikere, altså datautdannede folk, som sitter og vurder undersøkelsen – at den er gjort*

teknisk rett", og en person hos dem *"skal godkjenne hele undersøkelsen før den blir sendt tilbake til klinikerne, eller til røntgenavdelingen her"*.

Radiografen fortalte at *"vi på lab merker ikke noe til <DL-tjenesten>"*, bortsett fra dersom det har skjedd noe galt: *"Og så har det skjedd at det mangler noe i bildene, eller at de trengte en annen fase, eller noe annet – at det har vært en feil – da ser jeg også det. Da må vi sende noe nytt"* (2d). Denne korrespondansen var med en person knyttet til DL-tjenesten, og varierte med problemet som skulle løses.

4.2.3 Arbeidsflyt med DL-tjenesten

Den totale arbeidsflyten kan deles inn i fire trinn, som illustrert i Figur 21: (1) Ta CT-bilder, inkludert forarbeid (2) vurdere hvorvidt pasienten har sykdom som tilsier behov for invasiv undersøkelse (3) eventuell invasiv undersøkelse, og (4) eventuell behandling (på sykehuset). Tjenesten som studeres, *DL-tjenesten,* brukes for å avklare behov for invasiv undersøkelse for enkelte vanskelige tilfeller, og denne delen av arbeidsflyten vil derfor beskrives detaljert. I tillegg gjennomgår jeg det som skjer i trinnene før og etter, i den grad de påvirker eller blir påvirket av bruk av denne tjenesten.



Figur 21 Case 2: Hovedtrinn i arbeidsflyt for pasienter med mistanke om koronar sykdom

Forarbeid og billedtaking

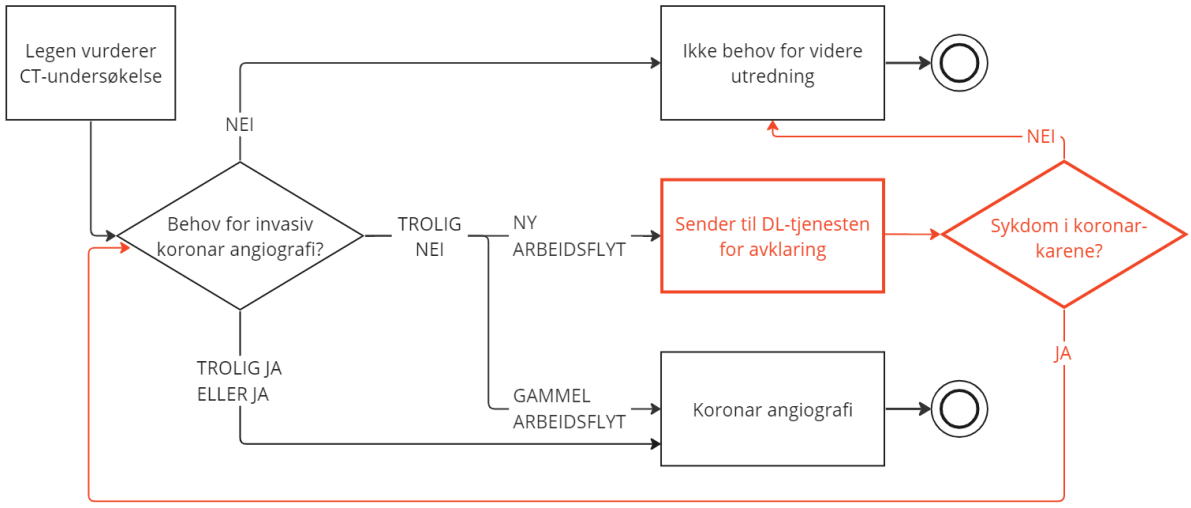
Det blir gjennomført en rekke tiltak for å få CT-bilder som er optimale for tolkning. Noen tiltak gjøres av pasientene selv, som at de ikke kan drikke kaffe eller røyke før undersøkelsen, men de fleste gjøres på sykehuset før selve billedtakingen. Også før man tok i bruk DL-tjenesten hadde man prosedyrer for å sikre at CT-bildene skulle være så enkle å vurdere som mulig. I tillegg er det gjort endringer i enkelte prosedyrer i forbindelse med innføringen, som å gi nitroglyserin i perfekt tid før

undersøkelsen, og ta bilde av større del av hjerteslaget for å få gode 3D-modeller av hjertet og kransarteriene.

Vurdering av behov for invasiv undersøkelse (koronar angiografi)

Det er røntgenlegen som ut fra bildene vurderer hvorvidt det er trangheter av funksjonell betydning i de store kransårene og dermed behov for koronar angiografi. Vurderingen gjøres i *"ett og ett kar av gangen".* Røntgenlegen fortalte at mange vurderinger oppleves enkle: *"Er det glatte kar, eller bare litt forandring? Det er enkelt [å vurdere]. Eller er det veldig trangt, så er det enkelt"*.

På et par områder er det imidlertid vanskeligere å vurdere om pasienten skal følges opp med koronar angiografi, eller om endringene er uten funksjonell betydning og pasienten derfor skal sendes hjem uten videre oppfølging fra sykehuset. Røntgenlegen fortalte at *"det er den lille gruppa som ligger rundt 50% [tetthet] som kan være vanskelig". "Hvis det er rett under 50 prosent – det er ikke så trangt at det skal behandles, men jeg er likevel litt usikker – de pasientene er det til nå jeg har brukt å sende til DL-tjenesten, for å få bekreftelse på at jeg har tenkt rett (..) Det andre er dersom det ikke nødvendigvis er trangt, men lengre område der det er sykdom i årene der vi er usikre på om den lettgradige trangheten totalt sett kan gi en reduksjon i blodstrømmen"*. Figur 22 viser en overordnet sammenligning av gammel og ny arbeidsflyt.



Figur 22 Case 2: Overordnet sammenligning av gammel og ny arbeidsflyt

Merk at figuren er forenklet for å tydeliggjøre forskjellene bedre. For eksempel er vurderingene i arbeidsflyten er mindre lineære enn figuren viser, og i enkelte tilfeller er det uklart om en undersøkelse bør sendes til DL-tjenesten eller ikke. Da drøftes undersøkelsen tverrfaglig med kardiolog, noe figuren ikke viser. Fullstendig arbeidsflyt er beskrevet i Figur 24.

Intervensjonskardiologen fortalte om vurderingene som blir gjort: "Det vi er ute etter svaret på her, er den trangheten som vi ser i blodårene. Har den betydning for blodstrømmen eller ikke? Er den det som vi kaller funksjonell betydning? Når vi gjør koronar angiografi, så kan vi se tranghet, måle, og så kan vi si at den har ingen betydning. Den har ingen funksjonell betydning. Men på CT-undersøkelse er det ren anatomi, og hvis du blir usikker, eller det er kalk som forstyrrer, eller det er i en slik grad at du ikke kan si at den har betydning eller ikke, så kan du ikke bare ut fra bildene si at den har funksjonell betydning. Så det svaret vi er ute etter er om det har funksjon, eller ikke. Det er beslutningen for dét, som vi er ute etter med kunstig intelligens" (ik).

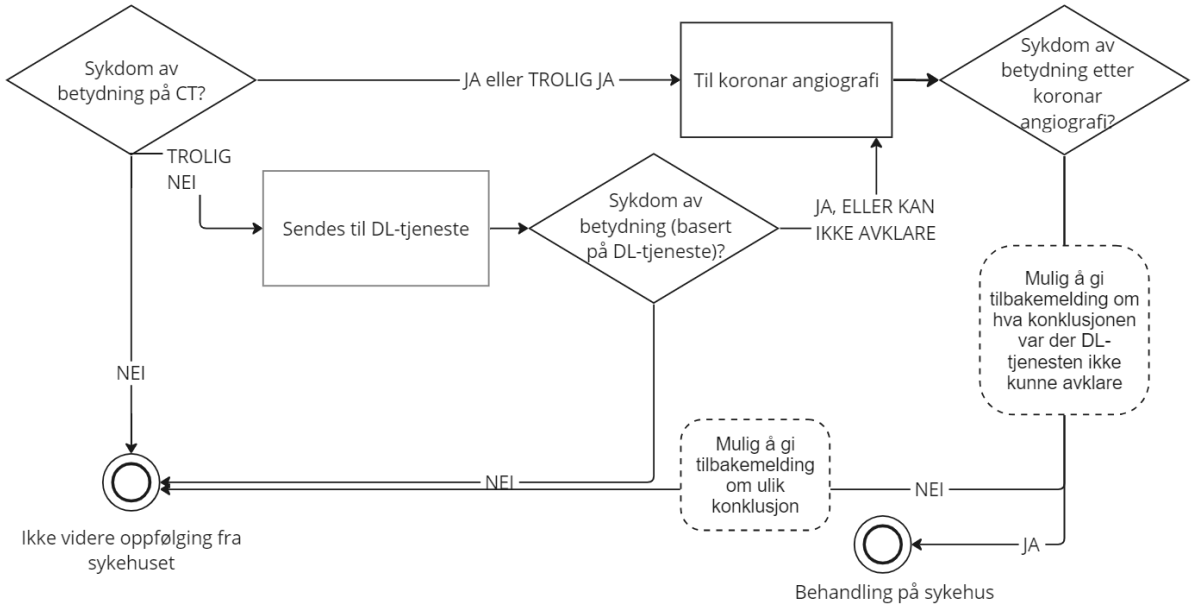
Bruk av tjenesten og videre pasientforløp

Dersom helsepersonellet bestemmer seg for å bruke DL-tjenesten, så er det "noen få tastetrykk og så er undersøkelsen sendt over" (rl). Etter innsending "så går det noen få timer bare, fra de er sendt inn til det kommer tilbake et resultat" (ik), "svartiden er fra under 4 timer til 6 timer" (rl). Deretter distribueres resultatet til de som måtte ha bruk for det.

Røntgenlegen fortalte at "noen ganger er det kanskje sånn at de hadde forventet at det skulle være noe, eller at det kan være litt motstridende svar – da ser vi på bildene sammen" (rl). "Jeg vurderer bildene kun på bakgrunn av hva jeg ser, og det beskriver jeg, og så må de [kardiologene] ta klinikken inn i det her". "På invasiv angio kan de også være i tvil om det er en stenose som bør behandles (…) – få en stent eller ikke. Er de usikre på det, så legger de inn et kateter, som heter FFR-kateter, og som måler trykket før og etter stenosen" for å vurdere hvorvidt det skal settes inn stent eller ikke. Dersom pasienten sendes videre til invasiv undersøkelse hos kardiolog, vil røntgenslegens svar på CT-undersøkelsen ses i sammenheng med andre data om pasienten: "Alle mine røntgensvar blir også vurdert av kardiolog, fordi de sitter jo med pasienthistorien", fortalte røntgenlegen, "og det er jo den invasive kardiologen som må vurdere sånne undersøkelser: Henger det sammen med klinikken? Kan vi gjøre noe med det? Er det noe forandring som vi kan sette inn stent på"?

Røntgenlegen fortalte at leverandøren ønsker tilbakemeldinger på vurderingene fra DL-tjenesten, for "da kan modellen deres lære og bli bedre", men det har ikke vært enkelt å få det inn i den daglige arbeidsflyten. "Det er litt vanskelig å følge det opp, for de [pasientene] går til invasiv undersøkelse

mange uker etterpå[32], og da må jeg lete opp pasienten og sjekke når det er den har fått time (…) det blir bare eget initiativ, og når det er travelt så følges ikke det godt nok opp". Resultatet fra invasiv angiografi er, ifølge Intervensjonskardiolog, "det vi regner som fasit". Siden denne bare blir gjort på pasienter med koronarsykdom eller mistanke om koronarsykdom, så er det bare en liten del av pasientene man kan gi tilbakemeldinger til leverandøren om. Dette er illustrert i Figur 23.



Figur 23 Case 2: Mulighet for tilbakemeldinger til leverandør

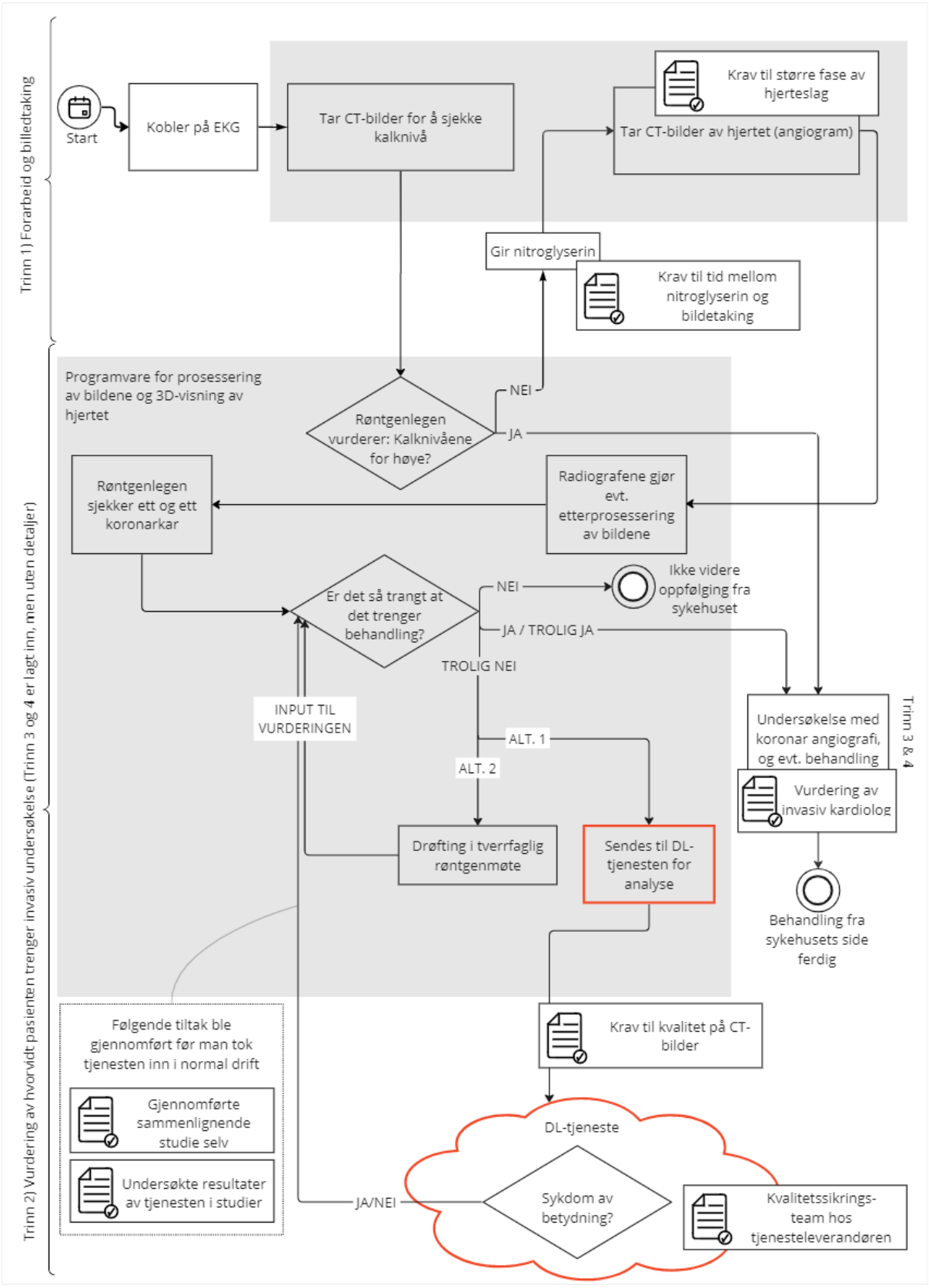
Tilbakemeldingsmuligheter til leverandøren er tegnet inn med stiplede kanter. Merk at pasienter uten sykdom av betydning sendes rett hjem uten verken avklaring hos DL-tjenesten eller invasiv undersøkelse, noe som gjør at man ikke har mulighet til å gi leverandøren tilbakemeldinger for disse pasientene. Figuren er en forenkling for å få frem muligheter for tilbakemeldinger. For en korrekt beskrivelse av selve arbeidsflyten, se Figur 24.

Overordnet flytdiagram for den nye arbeidsprosessen

Figur 24 viser overordnet arbeidsflyt ved bruk av DL-tjenesten. Trinnene har samme numre som i Figur 21. Trinn 1 er forarbeid og billedtaking. Merk den tette sammenflettingen av teknologi og medisin i dette trinnet: Kalknivå testes for å vite om det må gjøres manuell undersøkelse uansett,

[32] "etterpå" viser her til etter CT-undersøkelsen

eller om CT-bilder kan brukes til å avklare pasienten. Nitroglyserin gis i perfekt tid for å få åpne blodårer som gir bilder som er lette å tolke – riktignok også for mennesker, men rutinen ble endret for å få gode bilder til DL-tjenesten. Krav til bilde av større del av hjerteslaget gjør at det blir lettere å lage en god 3D-modell av hjertet. Tett sammenfletting (eng. imbrication) er kjent fra litteraturen og nevnt allerede i 2011 av Leonardi (2011), men det er likevel interessant å se hvor tett mennesker, medisinske intervensjoner og teknologi er sammenflettet. Neste trinn, trinn 2, skjer på programvare hvor de prosesserer CT-bildene og ser på 3D-visning av hjertet. Her er det bruk av DL-tjenesten er innført. Forskjellen er en, kanskje to, nye beslutninger. Akkurat som i case 1 vurderer de hvorvidt DL-tjenesten skal brukes. Når svaret kommer tilbake så er det for så vidt også i case 2 en beslutning på hvorvidt svaret fra DL-tjenesten skal hensyntas, men siden dette er et ekspertsystem som sjelden tar feil vil det som regel tas i betraktning. Dersom det er grunn til å gå videre med undersøkelser gjennomføres trinn 3, invasiv undersøkelse. Dersom man i trinn 3 finner sykdom av funksjonell betydning gjennomføres trinn 4, behandling.



Figur 24 Case 2: Oversikt over arbeidsflyt inkl. viktige kvalitetssikringstiltak

Figuren over viser oversikt over arbeidsflyt inkludert viktige kvalitetssikringstiltak ved bruk av DL-tjenesten. Grå bakgrunn viser hvilke trinn som gjøres på ulike maskiner/ i ulike programmer. Bruk av DL-tjenesten er markert med litt tykkere, rød ramme (i midten, langt nede). Vertikal tekst ytterst på venstre og høyre side, refererer til trinn-numre fra Figur 24.

4.2.4 Rammer for bruk

Etter anskaffelsen av tjenesten er det, som nevnt over, to hovedbeslutninger relatert til hvilket ansvar som skal gis til DL-tjenesten: Hvorvidt den skal brukes på en pasient, og i hvilken grad resultatet fra tjenesten skal brukes. Her vil jeg beskrive ulike rammer for bruk som ble fremhevet av intervjupersonene og som er knyttet til disse to beslutningene.

CT-bilder må være innenfor rammene for tjenesten

Leverandøren har en rekke krav til CT-bildene som skal sendes inn, og som dermed blir førende for om DL-tjenesten skal brukes (hovedbeslutning 1). Disse er oppsummert i Tabell 4. Noen av disse kravene var allerede oppfylt, mens andre krevde tiltak som ble satt i gang ved innføringen av DL-tjenesten. Merk at mange av disse kravene har medisinske konsekvenser og er derfor gode eksempler på at medisin og teknologi er tett sammenvevd.

Mulig utfordring med CT-bildene	**Tilpasning**
Høy (over 60 slag/min) eller uregelmessig puls gir mer bevegelse / uskarpe bilder	• Ikke røyk eller kaffe på undersøkelsesdagen • Får betablokkere på sykehuset
Vanskelig å vurdere tetthet i karene på CT-bildene, dersom karene ikke er tilstrekkelig utvidet	Nitroglyserin gis nå nøyaktig 4 minutter før undersøkelsen, etter tilbakemeldinger fra leverandør: *"Vi ser på klokka etter at vi har sprayet under tunga, og så etter fire minutter så skal vi starte"* (ir)

Ikke alle deler av bildet er skarpt på samme tid (endrer seg gjennom fasene av hjerteslaget)	Endring ble gjort etter tilbakemeldinger fra leverandør: "*Nå tar vi en litt større fase[33] av hjerteslaget, for å ha litt mer å gå på*" (rl)
Håndterer ikke ringformet kalkavleiring (kalk er generelt vanskelig)	*"Det skal ikke være kalk i karene som er som en ring"* (rl)
Andre faktorer: • *"Det skal ikke være for mange stenter[34] i karene, for eksempel, men den gruppa har vi ikke, uansett"* (rl). • *"De har sine føringer for hvilke snitt vi skal bruke, og hvilken kernel, hvor mange bilder og hvor stort field of view[35] og slikt"* (ir), men dette medførte ikke endringer, siden *"det var stort sett det vi brukte fra før av"* (ir)	

Tabell 4 Case 2: Oppsummering av utfordringer og tilpasninger for optimal bildekvalitet

Hastegrad og økonomi

Det er helsepersonellet som avgjør hvorvidt tjenesten skal brukes, og vedkommende vurderer dette i hvert enkelt tilfelle. Da vurderes blant annet hastegrad og økonomi.

Hvor raskt henholdsvis helsepersonellet og DL-tjenesten kunne analysere bildene, var relevant for hvilke tilfeller som kunne delegeres til tjenesten. Intervensjonskardiologen anslo svartiden til *"noen få timer, bare, fra de er sendt inn til det kommer tilbake et resultat".* Det gjorde at helsepersonellet kunne få en ekspertuttalelse relativt raskt, men ikke raskt nok til akutte tilfeller. Intervensjonskardiolog fortalte om denne pasientgruppen: *"Dette var jo pasienter som var til utredning for kransåresykdom, og som hadde lav eller moderat risiko for det. Stabile pasienter, ikke pasienter med hjerteinfarkt eller som lå på sykehus, men som var hjemme og hadde forskjellige typer plager som gjorde at de ble utredet for kransåresykdom"*.

[34] Stent er et lite rør som legges inn i blodårene for å holde det åpent
[35] Field of view (FOV) er området man tar bilde av, og som vises på skjermen i programvaren de bruker

Røntgenlegen fortalte at det var økonomiske rammer for bruk av DL-tjenesten, *"mest på grunn av prisen (…) sånn som jeg tenker nå at – hvorfor skal jeg sende det avgårde hvis de skal på invasiv angio uansett*"? Intervjupersonene kjente ikke til de økonomiske beregningene bak bruken, *"men det er i hvert fall dyrt å få pasientene fløyet tilbake til <Sykehusnavn>, og så er det en natt på sykehushotellet og så er det en behandling på lab"* (rl), *"så de beløpene blir ofte større enn det produktet koster, og da har vi selvfølgelig ikke heller regnet med det samfunnsøkonomiske (…), men det vil være motivasjonen for helseforetaket - i tillegg til bedre medisinsk behandling, da"* (ik).

4.2.5 Erfaringer fra innføringen

God kvalitet

Røntgenlegen fortalte at øynene til mennesker i noen tilfeller kan være en begrensning når det gjelder tolkning av bilder, noe som gjør at det er hensiktsmessig å delegere oppgaven til DL-tjenesten: *"Øynene ser veldig bra, men det er et punkt hvor du da ikke helt vet… heller det til den ene eller andre siden i forhold til om det er 50% trangt".* Han fortalte videre at DL-tjenesten gjennomfører *"en måling på hvor trangt det blodkaret er, som vi ikke klarer å gjøre med det blotte øyet"*.

Siden helsepersonellet bruker DL-tjenesten for å kvalitetssikre eller vippe egne svar, er det behov for høy kvalitet på vurderingene fra DL-tjenesten, ellers vil ikke rådene være til hjelp. *"Det betyr ganske mye for oss. Vi har sett at det er god kvalitet på produktet, sånn at det stoler vi ganske mye på, det resultatet der. Så det er med på å hjelpe oss å beslutte, skal vi tilby pasienten en koronar angiografi, altså en invasiv undersøkelse, eller skal vi si til pasienten at det er helt trygt å leve med, og du bør ha medisiner for at det ikke skal bli verre, men du trenger ikke å komme for en avklaringsundersøkelse hos oss. Så det betyr veldig mye i beslutningsprosessen hos oss hos de pasientene"* (ik).

Røntgenlegen fortalte at beslutningene ble bedre*: "Jeg synes at jeg kan ta en bedre avgjørelse, når jeg har <tjenestenavn>-analysen, og jeg kan trekke den inn i beslutningen på hva jeg mener svaret skal være. En slags … kollega … men man kan jo ikke stole blindt på det her (…), det er flere faktorer å vurdere. Men jeg synes jeg kan gi et bedre svar når jeg har den i bakgrunnen"* (rl).

Røntgenlegen fortalte at bruk av DL-tjenesten gav ham ekstra trygghet: *"Så sånn sett så vil det gi meg tryggheten av at når jeg tenker at det er det [behov for koronar angiografi], så er det det. Det*

jeg har tenkt er negativt[36], det kan <leverandørnavn> bekrefte. Det hadde vært verre at jeg tenkte at dette her er negativt, og at det til stadig vekk kom at det er positivt fra <navn på leverandør>".

Intervensjonskardiolog fortalte at det er lite uenighet knyttet til konklusjonen til DL-tjenesten, og at den hjelper personellet med beslutninger: *"Vår radiolog eller røntgenlege har jo vurdert at vi ikke kan avklare på grunnlag av CT-bildene, og i de aller fleste tilfellene så vil dette produktet kunne avklare det likevel. Så sånn sett så er det jo to forskjellige vurderinger – det er jo derfor vi bruker det her produktet. Men [det er] så å si aldri slik at produktet sier at det her er negativt resultat, og så kommer vi etterpå og sier at det er positivt likevel. Og heller ikke omvendt at produktet sier at det er positivt resultat og vi sier at det er negativt"*.

Røntgenlegen fortalte at *"det de sier er vel at de menneskelige øynene er ikke like gode som datamaskinen"* og *"det har hendt at trangheten har vært mye trangere enn det jeg har sett"* eller at *"jeg har ment at det var veldig trangt, og de [DL-tjenesten] har ikke påvist den trangheten så trang som jeg har ment"* (rl). Inn imellom har også helsepersonellet fått tilbake falske positive svar. Røntgenlegen fortalte om dette: *"Noen ganger har jeg lurt… jeg har tenkt at; Var bildekvaliteten bra nok? Er det bakgrunnen for at det blir litt [feil] (…) Det har vært noen få ganger, det begynner å bli en stund siden, hvor jeg mener at det har vært litengrann bevegelse i bildet, og jeg kan se det på CT-bildene, og da har det vært stenose hos dem [DL-tjenesten] som jeg kanskje ikke har vært enig i".*

Røntgenlegen fortalte videre at grenseverdiene i det DL-baserte produktet er satt *"sånn at de får noen falske positive fra <produktnavn>, altså de tolker det som stenose, og så kommer det på koronarlaben og så er ikke det stenose som vi skal stente – men det er for at vi ikke skal miste noen. Si at 'dette er ikke noe' – og så var det galt! De har lagt lista si sånn at man fanger opp litt flere falske positive stenoser".*

Sparer pasienter for invasiv undersøkelse

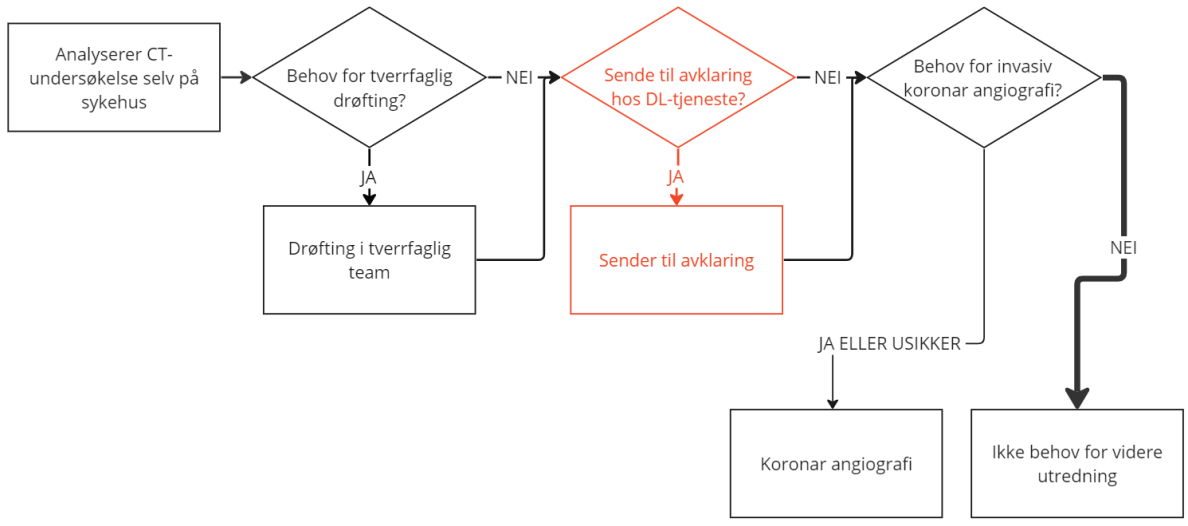
Intervensjonskardiolog fortalte at en av motivasjonsfaktorene var å spare pasienter for invasiv undersøkelse: *"Det som var motivasjon og det som vi så resultatet etter i innkjøringsperioden (…) var jo at andelen pasienter med koronar angiografi der det ikke førte til intervensjon – der det var ingen trangheter som trengte behandling – den gikk jo markant ned. Så det var færre pasienter som vi*

[36] At det "er negativt" betyr "negativt svar", altså at vedkommende ikke er syk, og derfor ikke trenger invasiv undersøkelse (som er positivt for pasienten)

undersøkte som hadde negativt funn, som ikke trengte undersøkelser". Før man tok i bruk DL-tjenesten måtte alle likevel ta koronar angiografi, der man går inn i blodårene og gjør en undersøkelse i for å bekrefte eller avkrefte sykdom. Ved å bruke DL-tjenesten kan helsepersonellet nå skille ut flere som kan slippe denne undersøkelsen. Videre fortalte han at andelen pasienter som nå ble avklart uten invasiv undersøkelse var stor, selv om han ikke har tall på det: *"Det tør jeg ikke si helt sikkert, for de tallene har jeg ikke. Av de planlagte koronar angiografi så lå vi vel tett opp mot… ja… om det ikke var halvparten eller over halvparten før vi innførte CT? Og så droppet jo det da vi innførte CT, men så har det droppet ytterligere når vi har innført <DL-tjenesten> også"*.

Trolig ikke dyrere

Denne studien har ikke undersøkt utgifter og økonomiske gevinster, men røntgenlegen fortalte om kostnader og økonomiske gevinster at *"rapporter beskriver det at det kan være opp i… jeg tror rundt tretti prosent besparelse på å bruke DL-tjenesten, fordi de får ikke unødvendig koronarangiografi og så videre, men om utenlandske rapporter gjelder for Norge, det vet jeg ikke. Men det er i hvert fall dyrt å få pasientene fløyet tilbake til <sykehusnavn>, og så er det en natt på sykehushotellet og så er det en behandling på lab – og hvis det viser seg at den er negativ[37]… Så sånn sett vil jeg tenke at helseforetaket sparer noen penger"*, i tillegg er det *"pasientvennlig og det synes jeg er det viktigste"* (rl).



Figur 25 Case 2: Flere pasienter kan avklares uten invasiv undersøkelse

[37] Negativ betyr her at det ikke er sykdom, altså var undersøkelsen strengt tatt unødig

Tykkelsen på pilen "Nei" til høyre skal vise at det er flere som nå havner i denne kategorien uten å måtte ta en invasiv undersøkelse. Merk at figuren er forenklet for å visualisere ett av målene med å ta i bruk DL-tjenesten. Fullstendig arbeidsflyt er beskrevet i Figur 24.

Tilpasningsbehovene opplevdes som små

Tilpasningsbehovene blir av helsepersonellet oppfattet som små. Intervensjonskardiolog fortalte at tilpasningene stort sett var *"det som røntgenlegen gjør med å sende bildene til <produktnavn>. Ellers så er det ikke noe særlig med tilpasninger som er gjort. Altså, vi, har jo tilpasset oss så vi kan bruke det her som en del av beslutningsgrunnlaget, men det er ikke gjort noen endringer på… på måten vi selekterer pasienter til CT-undersøkelser, eller måten vi gjennomfører pasientflyt eller sånne ting"*.

Røntgenlegen fortalte at *"jeg synes det fungerer fint, jeg, men jeg gjør jo arbeidet med å klikke inn alt selv, det er ingen som kommer og legger ting inn på pulten til meg, men i og med at det er så få pasienter, så synes jeg det her fungerer kjempefint"*. Det har imidlertid ikke vært like enkelt hele tiden, fortalte han videre, og i en periode ble derfor oppgaven gitt til kontorpersonale. I begynnelsen var det *"en masse, masse steg for å sende bildene avgårde, og så var det noen sekretærer som overtok den oppgaven, som vi hadde et fint system på, men nå har jeg tatt den over selv – det er så enkelt"*, fortalte han.

4.2.6 Oppsummering

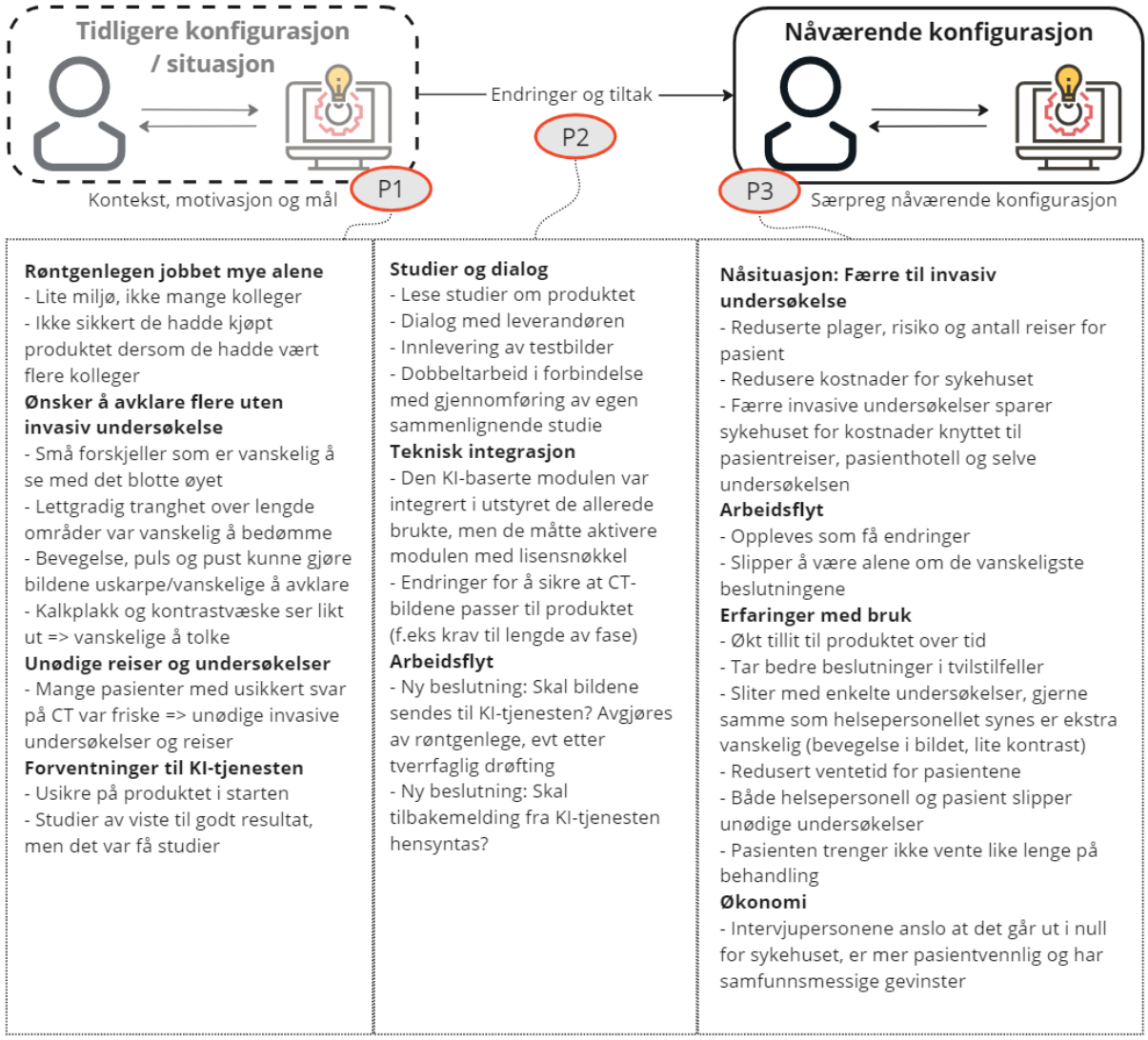
Enhet 2 bruker en DL-tjeneste for diagnostisering av pasienter med mistanke om kransåresykdom. Arbeidet ble igangsatt fordi mange som sendes til invasiv undersøkelse ikke er syke. Røntgenlegen arbeider alene, og har nytte av en tjeneste for avklaring av enkelte tvilstilfeller. Sykehuset kan spare tid og penger på å avklare disse uten invasiv undersøkelse, og pasientene slipper unødig reisetid og undersøkelse.

Produktet er trent på noen titalls tusen CT-undersøkelser, ifølge leverandørens nettsted. Det var relativt lite forskning på resultater fra tjenesten, så sykehuset gjennomførte en testperiode hvor de sammenlignet invasive undersøkelser med resultater fra tjenesten, og undersøkelsen viste at tjenesten gav gode resultater.

Sykehuset har gjennomført en anskaffelse, og det tekniske implementeringsarbeidet har vært lite, siden produktet er integrert i programvare som de allerede bruker. Før de tok produktet i daglig bruk, gjennomførte de en komparativ studie mellom resultater fra DL-studien og invasiv undersøkelse. Siden sykehuset betaler for hver enkelt bildeserie de sender inn, må helsepersonellet for hver aktuell pasient gjøre en vurdering av økonomien i å sende inn bildene til DL-tjenesten.

Helsepersonellet er svært fornøyde med at tjenesten gir svar av høy kvalitet innenfor rimelig tid, er enkel å bruke, og har redusert andelen pasienter som går til invasiv undersøkelse. DL-tjenesten har enkelte svakheter når bildene ikke er av helt topp kvalitet, men dette skjer sjelden siden det er krav til god teknisk kvalitet på innsendte undersøkelser.

Figur 26 oppsummerer hovedtrekk knyttet til de tre delspørsmålene for case 2, også utover de som jeg har skrevet mer utfyllende om i teksten over.



Figur 26 Case 2: Oppsummering av delproblemstilling 1-3

5 Diskusjon

I dette kapittelet diskuterer jeg funnene opp mot eksisterende forskningslitteratur, og beskriver implikasjoner for praksis og forslag til videre forskning. Jeg har plukket ut det som jeg mener er de mest interessante funnene innenfor hver delproblemstilling – for eksempel fordi det var overraskende, så ut til å være spesielt relevant for sosiotekniske konfigurasjoner med dyp læring, eller fordi jeg så konturene av mulige konsepter. Jeg undersøkte om det var forsket på noe lignende, og om det samme var funnet før. Siden dette er en sammenlignende studie, så har jeg vektlagt å belyse både fellestrekk og ulikheter for hver av delproblemstillingene.

I 5.1 forteller jeg om konteksten for ønske om endringer, som kort kan sammenfattes med behov for å håndtere ressursutfordringer, ønske om å forbedre kvaliteten på tjenestene og ønske om å redusere belastninger for pasient. 5.2 handler om endringer som ble gjennomført, både før og under innføringen, og i selve arbeidsflyten, og hvordan disse endringene kan se ut til å henge sammen med DL-systemenes uløselige forhold til data. I 5.3 tar jeg for meg den siste delproblemstillingen – hva den nye konfigurasjonen er kjennetegnet av. Jeg starter med å gå inn i detaljene og beskriver forslag til konsept for ulike tilnærminger til bruk av kunstig intelligens-systemer basert på funnene i de to casene. Deretter zoomer jeg ut igjen og ser på likheter knyttet til bruk av systemer basert på dyp læring som en del av ekspertbeslutninger. Jeg avrunder kapittelet med 5.4, som omhandler implikasjoner for praksis og for videre forskning.

5.1 Ressursutfordringer var sentralt for ønske om innføring av KI-system

Jeg spurte intervjupersonene om hva som var motivasjonen for, og målet med å ta i bruk DL[38]-systemene. Her sammenfatter jeg forskjeller og likheter. Det er sannsynlig at det var flere små og store faktorer som bidro til motivasjon for å ta i bruk DL-systemet enn de som oppsummeres her, og

[38] DL = dyp læring

at personer i andre roller enn de jeg intervjuet hadde andre mål og motivasjonsfaktorer for denne innføringen enn de som jeg intervjuet.

Ikke overraskende var det å håndtere kontekstspesifikke ressursutfordringer knyttet til eksempelvis eldrebølgen og en presset situasjon med tanke på tilgang til helsepersonell viktige motivasjonsfaktorer for de ansatte, og derfor også en viktig del av konteksten for ønske om endringer. Spesielt tydelig var dette i Sykehus 1 hvor helsepersonellet ønsket å posisjonere seg for utfordringene knyttet til en aldrende befolkning og flere pasienter uten tilsvarende økning i ressurser. De hadde behov for noe som kunne frigjøre kapasitet hos de ansatte til andre oppgaver, noe som er en kjent motivasjonsfaktor for bruk av kunstig intelligens fra litteraturen (Baird & Maruping, 2021; Nedelkoska & Quintini, 2018). I Sykehus 2 var situasjonen før de tok i bruk DL-tjenesten, at ca halvparten av såkalte *invasive*[39] undersøkelser kunne vært unngått. Å redusere antall unødige undersøkelser ville derfor kunne gi både kapasitetsmessige og økonomiske besparelser for sykehuset.

I et langstrakt land som Norge, med 19 offentlige sykehus (Helsepersonellkommisjonen, 2023) som skal bemannes, er det forståelig at ikke alle sykehus kan bemanne alle enheter med et så stort antall ansatte som man kanskje skulle ønske. Et interessant aspekt ved bruk av DL-tjenesten i Sykehus 2 var derfor at sykehuset brukte den for å bøte på ulempene ved å være et lite fagmiljø med litt få diskusjonspartnere. I dette tilfellet var det kanskje litt lite ressurser på røntgen, noe som røntgenlegen kjente ekstra godt når det var vanskelige caser og han manglet noen å drøfte med. Kapasiteten var kanskje likevel ikke så liten i det daglige at det ville være økonomisk ansvarlig å bemanne med en person til.

Et annet fellestrekk mellom Sykehus 1 og Sykehus 2 var ønske om å forbedre kvaliteten – riktignok med litt ulike inngangsvinkler. Sykehus 2 var, som allerede nevnt, et lite miljø hvor røntgenlegen ikke hadde mange kolleger å drøfte tvilstilfeller med. Å få støtte fra et DL-system kan ifølge teorien gi bedre kvalitet på beslutningene (Seeber et al., 2020). Også i Sykehus 1 tok de ansatte opp kvalitetsheving som en motivasjonsfaktor. I et kort tidsperspektiv så de en mulig, men liten kvalitetsgevinst i at DL-modulen kunne tegne forslag til segmentering for dem, noe som kunne redusere uønsket variasjon. For brystkreftpasienter ville imidlertid ikke en slik endring gi vesentlig

[39] *Invasiv* betyr i denne konteksten at man fører instrumenter inn i blodårene for å undersøke dem innenfra

kvalitetsgevinst, siden bilder av bryst oftest endrer seg relativt lite og sakte, og bildene i dag uansett bare tas en gang for hver pasient. Den største kvalitetsgevinsten mente de derimot ville komme på sikt, når de en gang tar i bruk segmentering før *hver* strålebehandling for å få en helt nøyaktig behandlingsplan hver gang. Dette vil i enda større grad være aktuelt når de tar i bruk KI på krefttyper der anatomien på bildene endrer seg mer og raskt fra gang til gang, som ved prostatakreft.

Det kan være greit å ha i bakhodet at helsetjenester i Norge er strengt lovregulert, og selv om det blir et større innslag av automatisering og autonome verktøy innenfor helse, er det *"fortsatt helsepersonellets ansvar at helsehjelpen som ytes er forsvarlig"*[40] (*Nasjonal strategi for kunstig intelligens*, 2020). I tillegg skal virksomheten arbeide *"systematisk for kvalitetsforbedring og pasientsikkerhet"*[41]. Ingen av intervjupersonene nevnte dette som årsak til deres fokus på kvalitet, men det er ikke usannsynlig at det er implisitt kunnskap som personellet har, og som derfor likevel har noe å si for motivasjonen for å sikre eller forbedre kvaliteten.

I Sykehus 2 ble reduksjon av unødige reiser og undersøkelser pekt på som en viktig årsak til bruk av DL-tjenesten. Å hensynta pasienter er empatiske handlinger som ifølge Rai (2019) mennesker er gode på å gjøre, og i henhold til Burton et al. (2020) er forventninger og følelser viktige faktorer som vektlegges når mennesker gjør vurderinger knyttet til hvorvidt de vil delegere til et system. Jeg regner med at denne handlingen ikke bare er knyttet til emosjonelle faktorer, men også rasjonelle faktorer som økonomi – som også ble nevnt, regelverk for helsetjenesten, profesjonsetikk og mer. Det er også kjent fra litteraturen at sosiale normer kan legge press på enkeltpersoner og påvirke delegeringsvilligheten (Burton et al., 2020), og eventuelle sosiale normer knyttet til det å fokusere på å spare pasientene for unødige reiser kan også ha vært en faktor inn i dette caset. I case 1 ble færre pasientreiser nevnt som en motivasjonsfaktor for en av informantene, men det ble samtidig oppgitt andre *hoved*motivasjonsfaktorer enn akkurat dette. En mulig tolkning av dette er at hensynet til pasientene i case 1 kan være å anse som en heldig sideeffekt som *forsterker* motivasjonen for å ta i bruk produktet, men som alene ikke er motivasjonsfaktor for bruken. Det er kanskje ikke overraskende at det var mindre fokus på reduksjon av pasient*belastninger* i Sykehus 1, siden produktet der ikke ble brukt til å beslutte videre forløp[42] for pasienten, men å få flere pasienter

[40] Dette er lovkrav i spesialisthelsetjenesteloven §2-2
[41] Dette er lovkrav i spesialisthelsetjenesteloven §3-4
[42] Forløp = hva som skal skje videre med pasienten

raskere til strålebehandling. Dette i motsetning til case 2 hvor bruken av DL-tjenesten påvirket videre pasientforløp ved å bidra i beslutning om hvorvidt pasienten måtte ta invasiv undersøkelse.

Seeber et al. (2020) viser til en rekke motsetningsforhold som kan gjøre seg gjeldende når man bruker kunstig intelligens inn i et team. Riktignok er disse motsetningsforholdene tenkt aktuelle når DL-systemene blir mer likestilte menneskene i teamene, men flere av dem tenker jeg er aktuelle å følge med på allerede i dag når virksomheter tar i bruk DL-systemer. Ett av motsetningsforholdene som Seeber et al. (2020) påpeker er at DL-systemer på den ene siden kan forbedre kvaliteten på beslutningene fra et team, men på den andre siden kan redusere den menneskelig evnen til å stille kritiske spørsmål. I Sykehus 1 har de besluttet at leger i spesialisering[43] (LIS) skal lære å segmentere manuelt før de får bruke segmenteringen generert av DL-modulen. Dette er et tiltak som kanskje kan redusere negative konsekvenser på menneskenes kapabiliteter til å stille kritiske spørsmål. Samtidig kan man få noe reduserte gevinster av bruken av DL-systemet, da studier har vist at det er nettopp de med mindre erfaring som kan ha størst nytte av kunstig intelligens (Rajpurkar et al., 2022).

Et annet av Seeber et al. (2020)s motsetningsforhold er at man tar i bruk DL-systemer for å utnytte fordelen av systemets høye tempo, men at dette i neste omgang kan føre til overbelastning hos teammedlemmene dersom de må kvalitetssikre alt fra et DL-system som aldri trenger pause. I denne undersøkelsen ble ikke dette nevnt som et problem, verken i case 1 eller 2. Det er nok flere grunner til det. En mulighet er at overbelastninger først blir plagsomme når de har vart en stund, slik at det kan bli en utfordring senere. DL-systemene brukes på få av pasientene i begge casene, og har derfor trolig lite å si for den totale belastningen på de ansatte: I case 1 brukes DL-modulen på en liten pasientgruppe, og i case 2 brukes DL-tjenesten kun på vanskelige tilfeller. En annen mulig forklaring er at de ansatte i begge sykehusene hadde god kontroll på hvor mye DL-systemet skulle brukes ved å sette det i gang selv. Dette vil teoretisk sett begrense gevinstene av DL-systemet, men har som fordel at systemet "går i takt" med de ansatte uten å overbelaste dem.

5.2 DL-systemenes forhold til sine treningsdata la føringer for endringsbehov og tiltak

[43] "Leger i spesialisering" (LIS) er leger som læres opp til å bli spesialister

Delproblemstilling 2 handler om hvilke endringer og tiltak intervjupersonene fortalte om i tilknytning til innføringen av DL-systemet i den nye konfigurasjonen. Det ble både gjort endringer *for* å innføre DL-systemene og det dukket opp endringer *som et resultat* av innføringen. Når det gjelder endringer knyttet til innføringen overrasket det meg at det var mye mindre teknisk integrasjonsarbeid på sykehussiden enn jeg forventet. Begge DL-systemene var eller ble integrert av leverandøren som en del av utstyret helsepersonellet allerede brukte. På begge sykehusene hadde de riktignok gjennomført noe integrasjonsarbeid, men i begge casene forklarte de at dette var av lite omfang.

I både Sykehus 1 og 2 hadde de gjort endringer eller tiltak av flere typer, både engangsoppgaver og endringer i rutiner, for å få en god integrasjon av DL-systemet. Det var påtagelig at tiltakene som ble nevnt i stor grad handlet om data, både kontroll på treningsdata og/eller inndata, samt vurdering av utdata (resultater). Ifølge (Robbins, 2020) er kontroll på data nødvendig for å la DL-systemer med redusert granskbarhet gjøre beslutninger. Det er ikke usannsynlig at disse tiltakene også bidro til at bruken av DL-modulen i case 1 opplevdes tryggere, selv om utfordringen der kanskje var mer knyttet til noe redusert *pålitelighet* heller enn redusert granskbarhet.

Det vakte min nysgjerrighet at det var stor heterogenitet i hvordan sykehusene hadde løst problemstillinger knyttet til data – som mengde og kjennskap til treningsdata, i hvilken grad og hvordan inndata ble tilpasset systemet, og i hvilken grad man hadde tillit til resultatet fra DL-systemet (utdata). I denne studien fikk jeg inntrykk av at grad av kjennskap til *treningsdata* fungerte som et veiskille for hvordan videre kvalitetssikring ble gjennomført, så jeg starter med å gjennomgå bruk av treningsdata i 5.2.1.

Jeg kunne ikke se vesentlig forskjell i hvor tilfredse de ansatte i case 1 og 2 virket med bruk av DL-systemet, til tross for flere ulikheter som jeg ville trodd kunne påvirket tilliten til det. Imidlertid hadde begge sykehusene god kontroll på inndata, og mulighet til å inspisere resultatet og gjøre seg opp en mening om det selv. Betydningen av kontroll på inn- og utdata diskuterer jeg derfor i 5.2.2.

I 5.2.3 drøfter jeg de ulike typene tiltak som jeg fant i denne studien, og foreslår å også studere DL-systemer i lys av litteratur som innehar et større element av uforutsigbarhet i seg enn litteratur knyttet til tradisjonelle informasjonssystemer.

5.2.1 Kjennskap til treningsdata for DL-systemer i medisin – fordel eller forutsetning?

Til forskjell fra andre IT-systemer, så kan ikke DL-systemer lages uten data. Tilgang til data for trening er derfor helt essensielt for å kunne lage systemet, og treningsdata er av avgjørende betydning for hvor godt systemet vil fungere (Ozkaya, 2020). Siden DL-systemer trenes og funksjonaliteten derfor ikke kan spesifiseres på samme måte som man gjør med tradisjonelle IT-systemer, introduseres det en usikkerhet knyttet til hvordan systemet vil fungere på andre data i produksjon (Ozkaya, 2020). Som en del av kvalitetssikringen er det derfor nødvendig å validere DL-systemet på sykehusets data (Makhlysheva et al., 2022). Det er fremdeles ikke mye erfaring knyttet til ekstern validering[44] (Rajpurkar et al., 2022), så det var interessant å høre de ansattes ulike vurderinger knyttet til kvalitetssikring av produktets ytelse.

Gitt DL-systemenes tette forhold til treningsdata, er det kanskje ikke så rart at Sykehus 1 valgte å samarbeide tett med sin leverandør og lage de treningsdataene som skulle brukes selv. Det er regnet som vanskelig å etablere korrekt "fasit" (*"ground truth"*) innenfor medisin (Asan et al., 2020), blant annet på grunn av høy usikkerhet, og ufullstendige, heterogene og unøyaktige data (Esteva et al., 2017), og helsepersonellet i case 1 brukte *"ganske mange timer"* på å lage treningsdata i form av korrekt segmenterte CT-bildeserier. Selv om alle som laget treningsdata brukte den europeiske standarden for inntegning, så var det likevel litt forskjell mellom inntegningene som de ulike legene gjorde, og behov for ytterligere harmonisering. Modellen ble trent på CT-bilder fra over 200 pasienter, hver av disse med en bildeserie på ca 150 bilder. Siden sykehusets ansatte hadde laget treningsdataene selv, visste de at inndata til DL-modulen ville være svært like treningsdataene, og de kunne være relativt trygge på at resultater fra DL-modulen ville gi korrekt segmentering for alle "standard" pasienter. Man kan derfor si at Sykehus 1s engasjement i å lage så presise og gode treningsdata som mulig, var en viktig del av deres kvalitetssikring, og at sykehusets kvalitetssikring derfor startet allerede under samarbeid med leverandøren om produktutvikling. Siden de kjente treningsdataene godt, og planla å kvalitetssikre all segmentering gjennomført av DL-systemet, vurderte de at en vellykket ekstern validering på 15 CT-bildeserier[45] var nok før de tok DL-modulen i bruk. Jeg la merke til at de ansatte i Sykehus 1 brukte sin detaljerte kjennskap om treningsdataene

[44] Ekstern validering er testing av DL-systemet på sykehusets egne data (Rajpurkar et al., 2022)
[45] En bildeserie er på ca 150 bilder

hyppig i forklaringer knyttet til DL-modulen, som å forklare styrker og svakheter ved systemet. Jeg tolker det som at kjennskap til treningsdata var viktig for deres forståelse av DL-modulen, og forstod det også slik at det gav dem trygghet i bruken av modulen.

Sykehus 2 var mindre tett på leverandøren, siden de gjennomførte en anskaffelse av et internasjonalt utbredt produkt som leverandøren hadde helhetsansvar for. Helsepersonellet var i all hovedsak ikke med på videre opplæring av systemet, men deltok i starten på månedlige møter med leverandøren, der hensikten var å øke kvaliteten på resultatene ved å gjennomgå undersøkelser med ulik konklusjon fra helsepersonellet og DL-tjenesten. En vesentlig del av sykehusets kvalitetssikring gikk ut på å gjennomføre en komparativ studie for å validere resultatene fra ML-tjenesten på sykehusets egne data. Det er forståelig at de følte behov for dette, siden mange algoritmer for bildebasert radiologisk diagnose har vist seg å få redusert ytelse og betydelig ytelsesnedgang på eksternt datasett (Yu et al., 2022).Til studien trengte de "*det vi regner som en fasit*" (ik) – ikke for å bidra med treningsdata slik tilfellet var i case 1, men for å selv kunne ettergå resultatene fra DL-tjenesten før de tok produktet i daglig bruk. Det var kanskje enda vanskeligere å etablere denne fasiten i case 2 enn i case 1, siden man ikke kunne finne korrekt konklusjon uten en invasiv undersøkelse hvor man går inn i blodårene til pasienten og undersøker blodgjennomstrømmingen i koronarkarene[46]. Dette gjør man vanligvis bare på pasienter som man mistenker er syke, noe som betyr at man i utgangspunktet har sikker fasit bare for de syke eller mulig syke, og ikke for de friske. For å få evidens også på at de friske ble riktig diagnostisert gjennomførte de invasiv undersøkelse på alle pasientene, også de som helsepersonellet trodde ikke hadde sykdom av funksjonell betydning, i en periode. På den måten fikk de validert at produktet gav tilfredsstillende resultater på sykehusets egne data, og at det var trygt å bruke. I tillegg til egen studie gjennomgikk de det som fantes av studier av DL-tjenesten – noe som ifølge intervensjonskardiologen ikke var så mye. De var imidlertid kjent med at sykehus i mange land brukte produktet, noe som kan gjøre en tryggere på egen bruk, og etter hvert fikk også helsepersonellet egne, gode erfaringer med produktet i bruk.

Ansatte i analyseteamet hos tjenesteleverandøren i case 2 er kanskje de nærmeste brukerne av DL-systemet som er i kjernen av DL-tjenesten Sykehus 2 bruker. Det er verdt å merke seg at disse potensielt kjenner DL-systemets treningsdata godt. Analyseteamet kvalitetssikret og rettet 3D-modellene som DL-systemet genererte for dem, og kan derfor utgjøre en parallell til de ansatte i

[46] Koronarkarene = blodårene som forsyner hjertet med blod

Sykehus 1 som også kjente sine treningsdata godt, og kvaltetssikret og rettet resultatene fra DL-systemet. Selv om dette ville vært interessant å vite mer om, så var det utenfor scope av denne studien å intervjue leverandørene, og mitt kjennskap til dette er derfor begrenset til åpen informasjon som leverandørene har lagt ut på nettet. Av denne grunn er det også uklart om hvorvidt kjennskap til treningsdata "bare" er en *fordel* for trygg bruk av DL-systemer i helsetjenesten, eller om man bør se på det som en *forutsetning.*

5.2.2 DL-systemene i bruk: God kontroll på inn- og utdata

Det vekket min nysgjerrighet at de ansatte ikke opplevde det som problematisk at DL-systemet innimellom ikke var hundre prosent pålitelig. For eksempel var det som regel behov for å korrigere enkelte snitt på hver undersøkelse. Det var ikke uventet at systemet gjorde feil, siden det er kjent fra litteraturen at DL-systemene både kan ha feil (Huo et al., 2022) og være knyttet til store usikkerheter (Ozkaya, 2020). Feil fra utstyr brukt i helsetjenesten kan imidlertid føre til direkte skade på mennesker (Habli et al., 2020), og jeg trodde derfor at dette skulle oppleves å stå i et motsetningsforhold til både pasientsikkerhet og evidensbasert medisin[47], og derfor redusere tilliten til systemet.

"Jeg stoler på den med forbehold", sa onkologen. Helsepersonellets tillit til DL-systemet er avgjørende for å ta det i bruk (Makhlysheva et al., 2022), og i innebygde systemer er denne tilliten drevet av transparens og pålitelighet (Glikson & Woolley, 2020). I case 1 var det kanskje redusert pålitelighet som var mest tydelig, men flere faktorer kan kompensere, og ikke engang *lav* pålitelighet fører derfor automatisk til lav tillit og manglende bruk (Glikson & Woolley, 2020). Da er det kanskje ikke uventet at de ansatte i sykehus 1 ikke lot seg stresse av det de tross alt regnet som små feil. Tillit er avhengig av hva systemet skal gjøre, og emosjonelle og kognitive faktorer (Glikson & Woolley, 2020), noe som vi også ser her: I dette tilfellet skal systemet utføre segmentering, som i utgangspunktet er tidkrevende for mennesker, men raskt og lett å sjekke i etterkant av at DL-modulen har gjort jobben. Siden det er store tidsmessige fordeler med bruken, og helsepersonellet uansett går over og sjekker segmenteringen manuelt, er det ikke utenkelig at det måtte vært ganske mange eller store feil før de ville fått betenkeligheter med å bruke systemet. I tillegg hadde de god

[47] Les om evidensbasert medisin i Engebretsen, E. B., Hilde;. (2023, 22. mai 2022). *Kunnskapsbasert medisin i Store medisinske leksikon på snl.no*. Retrieved 9. april 2023 from https://sml.snl.no/kunnskapsbasert_medisin

kjennskap til systemet og forståelse av hva som kunne dukke opp av eventuelle feil. Sistnevnte var litt overraskende, siden DL-systemer er kjent for å kunne gjøre *uventede* feil (The Lancet Digital, 2022). Det er flere mulige forklaringer på dette. Kanskje de hadde brukt DL-modulen for kort tid til at det hadde dukket opp noe uventet, eller kanskje de hadde opplevd uventede feil, men ikke fokuserte på dem siden det var så få. En mulighet er også at noen av feilene som de ikke lar seg overraske av i dag, ville vært overraskende dersom de ikke kjente treningsdataene. "*Vi vet*", sa en av de ansatte, og de visste mye. De visste at DL-modulen bare gjorde feil på et fåtalls bilder, hvilke pasienter som DL-modulen ikke ville fungere på, og ofte visste de både hvilke bilder feilene ville komme på og hvorfor. Når DL-modulen i tillegg erstattet et produkt som både gjorde mange og store feil, så er det ikke uventet at de få feilene ble betraktet som helt uproblematiske.

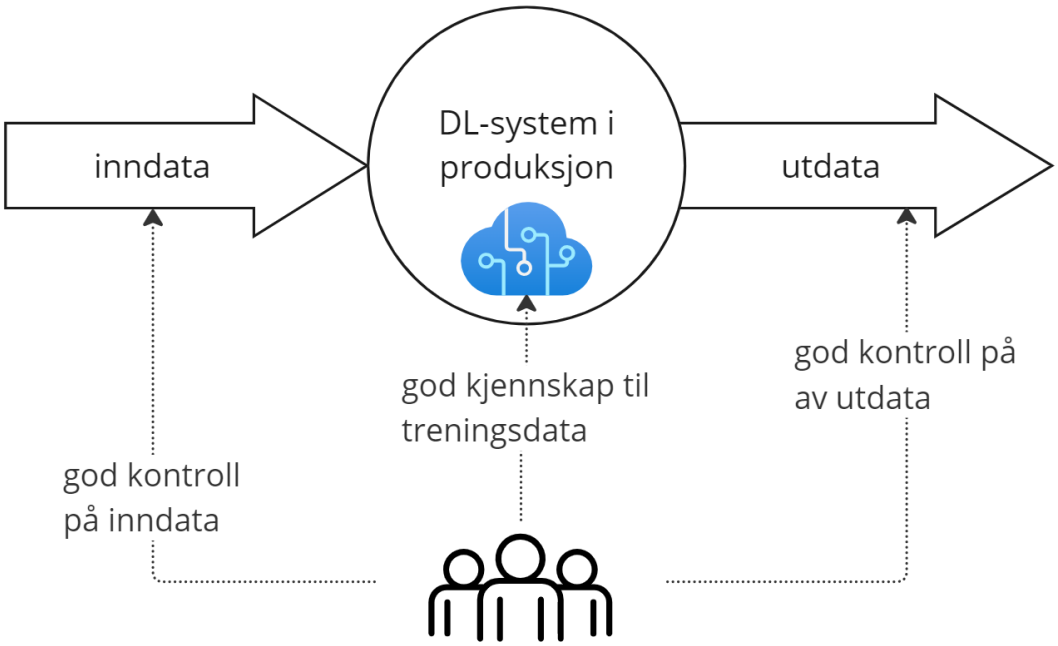
Som nevnt er leverandøren av case 2s DL-tjeneste holdt utenfor scope av denne studien, og jeg kan derfor ikke si noe om hvorvidt det er en parallell når det gjelder utfordringer og eventuell håndtering av upålitelighet fra selve DL-systemet. Imidlertid gjorde Sykehus 2 flere tiltak for å sikre riktige inndata av god kvalitet til DL-tjenesten. Tiltak som ble gjort i forbindelse med innføringen var for eksempel at helsepersonellet ble mer nøye på tiden fra man gav pasientene nitroglyserin til bildene ble tatt, slik at karene var så åpne som mulig under CT-scanningen, og bildene derfor ble lettere å tolke. De tok også bilde av en større fase av hjerteslaget, for å forsikre seg om at man fikk klare bilder av begge kransarteriene. Dette er kanskje et eksempel på situasjonen beskrevet i Coiera (2004), som mente at man ville oppleve at menneskene som i første omgang *skaper* sosiotekniske systemer i neste omgang *blir formet* av dem (Coiera, 2004). Om ikke helsepersonellet ble formet, ble i hvert fall helsepersonellets *meninger* om medisinske prosedyrer nok formet av DL-tjenestens behov, til at de forandret rutinene sine. Endringene i rutinene gjorde at CT-bildene ble av bedre kvalitet, og reduserte derfor sannsynligheten for feil fra DL-tjenesten.

For mange DL-systemer er det ikke mulig å skjønne systemet i detalj (Makhlysheva et al., 2022), og generelt kan det også være vanskelige å verifisere resultatene (Ozkaya, 2020). Det kan være noe lettere å kvalitetssikre DL-resultater basert på medisinske bilder enn enkelte andre datatyper, siden helsepersonellet har kompetanse til å vurdere de samme bildene som DL-systemet har vurdert, og dermed direkte vurdere hvor gode beslutningene er (Makhlysheva et al., 2022). Dette er i tråd med både case 1 og case 2. For eksempel kvalitetssikret de ansatte i Sykehus 1 inntegningen gjort av DL-modulen, og tegnet manuelt dersom noe var feil, noe som er et eksempel på tiltak som reduserte konsekvensen av eventuelle feil. Selv om det var sjelden at ansatte i Sykehus 2 var uenige med

resultatet fra DL-tjenesten, så hadde også legene der mulighet til å vurdere svaret fra DL-tjenesten kritisk. Røntgenlegen presiserte at man må bruke hodet og ta med flere faktorer i betraktning når man gjør beslutningen, og legene hadde alltid siste ordet uansett hva DL-tjenesten konkluderte med.

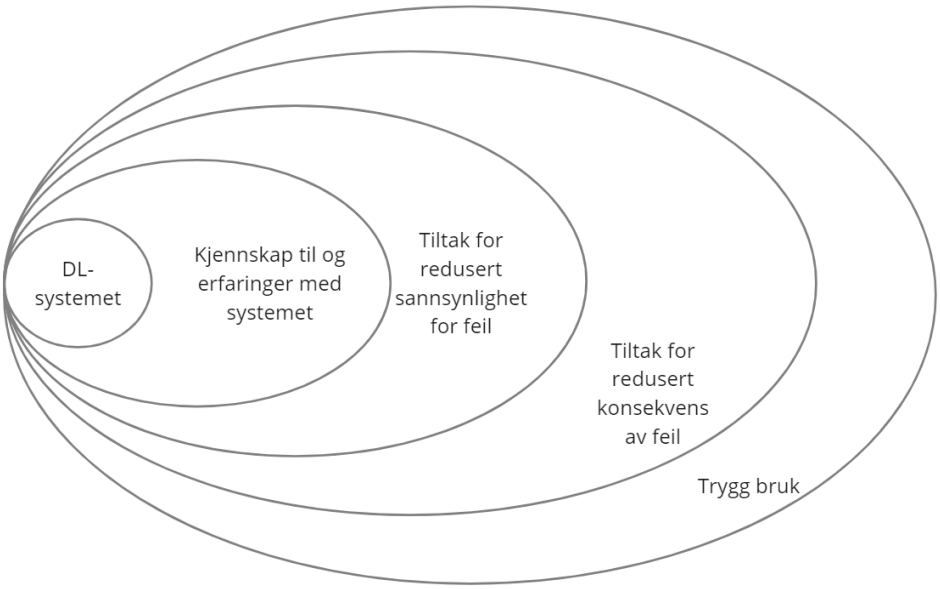
5.2.3 Uventede utfordringer reduseres ved kontroll på data

I hovedtrekk kan det se ut til at mange av endringene eller tiltakene som ble gjort i både case 1 og 2 handlet om (1) å ha kjennskap til eller kontroll på treningsdata, og (2) inn- og utdata (resultat), og at (3) god kontroll kan bidra til at de ansatte opplever bruken tryggere, som illustrert i Figur 27.



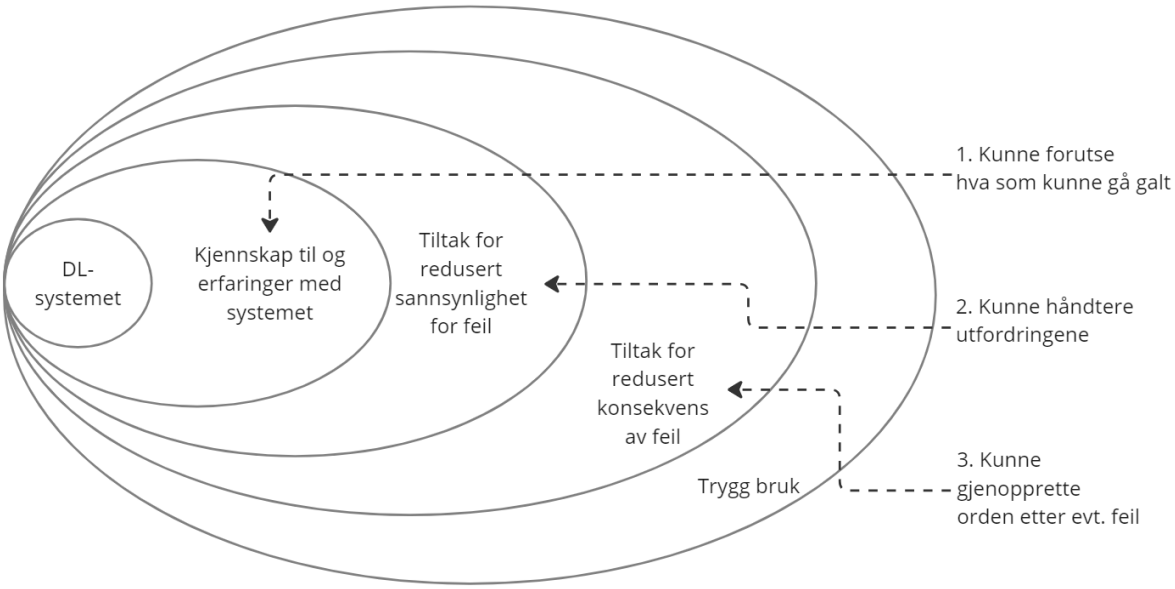
Figur 27 Summen av kjennskap til læringsdata og kontroll på inn-/utdata kan bidra til trygghet

Mange av tiltakene kan grupperes inn i tiltak for å redusere *sannsynlighet* for feil og *konsekvens* av feil. Figur 28 visualiserer funnene knyttet til dette: DL-systemet (i midten) kan oppleves tryggere i bruk ved at man (1) har kjennskap til og erfaringer med systemet, (2) basert på kjennskapen setter opp tiltak som reduserer sannsynligheten for feil og (3) setter opp tiltak som reduserer konsekvensene av feil.



Figur 28 Erfaringer og risikoreduserende tiltak rammer inn bruken av DL-systemet

I ML-systemer er usikkerhet en karakteristika (Ozkaya, 2020), beslutningene kan være vanskelig å forstå (Holzinger et al., 2019), og logikken for å komme frem til et svar er utviklet basert på prosessering av data (Ozkaya, 2020). Dette kan bidra til at eventuelle feil kan komme på uventet tid og være annerledes enn typiske feil begått av mennesker. Dette er en situasjon som vi kan gjenkjenne fra risikohåndtering og beredskapsarbeid, og nettopp dette gjør at det kan være relevant å se til litteratur fra pandemihåndteringen. I Figur 29 har jeg derfor utvidet Figur 28 ved å se den i sammenheng med Rangachari and L. Woods (2020)s forskning under pandemien, og deres prinsipper for hva som er viktig for å løse *uventede* utfordringer. Det kan fremstå som at det er en parallell her, og en mulig forståelse av funnet er derfor at de som arbeidet med DL-modulen opplevde at bruken var trygg fordi de både kunne *forutse* hva som kunne gå galt, hadde evne til å *håndtere* utfordringene og evne til å *gjenopprette* (Rangachari & L. Woods, 2020). I case 1 kunne man gjenopprette situasjonen ved å gjøre inntegningen på nytt, og i case 2 kunne man se bort fra svaret fra DL-tjenesten dersom det virket som at det ikke var riktig.



Figur 29 Tiltak for trygg bruk versus prinsipper for resiliente helsetjenesteorganisasjoner

Figuren viser eksempel på grupper av kompenserende tiltak som kan bidra til tryggere bruk. Gruppene er basert på fra denne studien, og er markert med sirkler til venstre i bildet. Disse kan ses i sammenheng med de tre prinsippene for håndtering av uventede utfordringer i Rangachari and L. Woods (2020)s rammeverk for resiliente helsetjenesteorganisasjoner. Disse er markert med løpenummer 1-3 til høyre i figuren.

5.3 Dyp lærings upålitelige natur preget konfigurasjonene

En konfigurasjon er et spesifikt sett av relasjoner mellom menneske(r) og maskin(er) med en gitt deling av oppgaver og ansvar mellom dem (Grønsund & Aanestad, 2020). Sykehus 1 og 2 hadde valgt å ta i bruk DL-systemene på forskjellig måte, og dette førte ikke bare til ulike typer endringer, men det preget også til en viss grad selve konfigurasjonen, noe som er tema for delproblemstilling 3. I den videre diskusjonen vil jeg starte med de store konseptuelle forskjellene i hvordan Sykehus 1 (avsnitt 5.3.1) og Sykehus 2 (avsnitt 5.3.2) har satt opp sine konfigurasjoner, før jeg diskuterer en likhet knyttet til disse to casene – at man lager et beslutningsstøttesystem for eksperter av noe som har en upålitelig kjerne (avsnitt 5.3.3).

5.3.1 Case 1: Personellet delegerer til en arbeidsvillig DL-assistent

I Sykehus 1 valgte de å la helsepersonellet og DL-modulen dele på ansvaret for segmentering. DL-modulens oppgave var å prosessere de store mengdene CT-bilder fra relativt standard pasienter, finne alle relevante strukturer, og tegne strek rundt disse (segmentere) for bruk i den videre planleggingen av strålebehandlingen. Dette var en tidkrevende oppgave for helsepersonellet, men tok bare ca et minutt for DL-modulen, ikke uventet siden DL-systemer kan være både raske og nøyaktige i bruk (Rai, 2019).

Personellets oppgave var å utføre tilsvarende arbeid på strukturer som DL-modulen ikke klarte å segmentere. DL-modulen var trent på "standard" pasienter, og hadde ikke teknisk forutsetning, og heller ikke nødvendig kognitiv fleksibilitet (Topol, 2019; Véras et al., 2015) til å håndtere de mer ustandard bildene. Helsepersonellet fikk i tillegg nye oppgaver. Noen var engangsoppgaver, som å lage og kvalitetssikre treningsdata under utviklingen, og andre var endringer i den daglige arbeidsflyten, som for eksempel beslutninger om hvilke pasienter DL-modulen skulle brukes på og hvorvidt DL-modulens segmentering skulle brukes. Helsepersonellets ansvar i denne sammenheng var å håndtere oppgavene knyttet til overvåkning og korrigering av feil segmentering fra DL-modulen, eller *feil respons i et ellers autonomt system*, som er ordlyden Rahwan (2018) bruker. De ansatte var den ansvarlige part, og gikk god for at segmenteringen hadde god nok kvalitet for videre bruk i arbeidsflyten. At helsepersonellet tok nettopp dette ansvaret er i tråd med Leonardi (2011) som sier at kvalitetskontroll er en oppgave som gjerne tas av mennesker.

Case 1s bruk av DL-modulen er et eksempel på human-in-the-loop-konfigurasjon, hvor mennesker og maskinlæring komplementere hverandre ved å stole på hverandres styrker og overvinne svakhetene (Teodorescu et al., 2021). Siden mennesket i denne konfigurasjonen har siste ordet og er den ansvarlige part, kan man også bruke begrepet human-in-charge (Kitamura, 2023). Fordelingen av arbeid og nye beslutninger som er innført, er på linje med Grønsund and Aanestad (2020) som sier at det er vanlig at menneske og maskin deler på ansvaret for en oppgave, og at nye oppgaver kan dukke opp når man tar i bruk kunstig intelligens. Mennesker og maskiner har ulik kapasitet og ulike kapabiliteter (Leonardi, 2011), og hensikten med å ta i bruk kunstig intelligens er gjerne å spare de ansatte for oppgaver som ML-systemet kan gjøre bedre (Asan et al., 2020; Canals & Heukamp, 2019; Rai, 2019; Topol, 2019), eller øke teamets ytelse ved å sette et ML-system sammen med ekspertene (Rajpurkar et al., 2022). I dette tilfellet klarte DL-modulen å gjøre nødvendige beregninger og inntegninger effektivt – på en brøkdel av det legene brukte, og som regel også nøyaktig, i hvert fall på standard pasienter. Effektivitetsgevinster er ikke uventet ut fra litteraturen

som sier at maskiner kan prosessere store datamengder raskt og med høy presisjon (Asan et al., 2020; Topol, 2019), være bedre enn mennesker i å gjenkjenne mønstre (Canals & Heukamp, 2019) og gi økt effektivitet innenfor radiologi (Langlotz, 2019).

Teodorescu et al. (2021) argumenterer for at rettferdigheten kan trues i DL-systemer, blant annet av statistiske sjeldenheter. I case 1 har man valgt å bruke fordelen helsepersonellet har fra årelang erfaring ved å la dem ta de mer ustandard tilfellene, noe som kan ses på som et risikoreduserende tiltak rettet nettopp mot statistiske sjeldenheter. I tillegg til dette kunne man se flere vanlige måter å omgå begrensninger i teknologi på, i helsepersonellets bruk av DL-modulen: De brukte ikke resultater fra DL-modulen når de ikke var bra nok, de delte arbeidsflyten opp i ett "spor" for CT-bilder som man kunne bruke produktet på og ett for resten av bildene, og de tilpasset kvalitetssikringen til den nye modulen. Dette er forskjellige, kjente måter å utøve fleksibilitet på, og beskrives på overordnet nivå i Leonardi (2011). Nå og da gjorde DL-modulen småfeil, noe de ansatte forventet og hadde sett et mønster i. De viet derfor ekstra tid til bildeserier der de forventet feil som måtte rettes. I tillegg brukte de tid på rutinemessig kvalitetssikring av øvrige CT-serier. Konfigurasjonen bar preg av at DL-modulen var en slags *arbeidsvillig assistent* for dem, som effektivt tok det store volumet av enkle oppgaver, men som ble fulgt tett opp siden den kunne gjøre feil, spesielt ved mer kompliserte oppgaver.

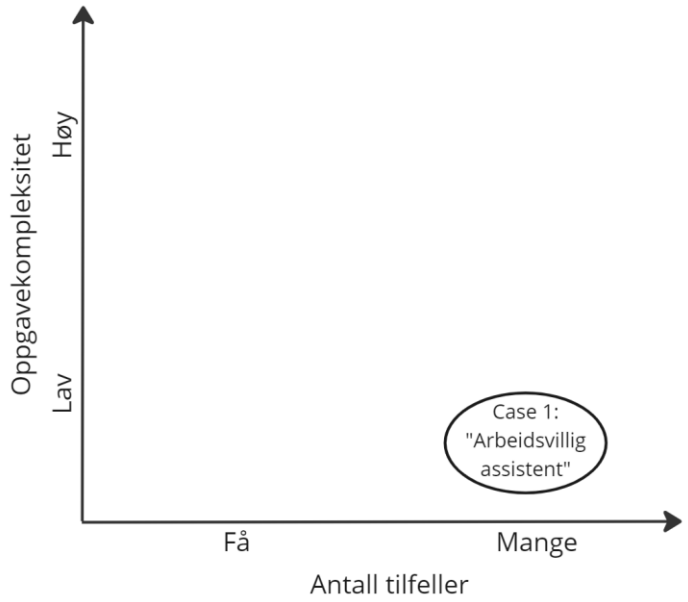
En viktig del av delegeringssituasjonen er en vurdering av hvorvidt en oppgave er egnet for delegering, noe som kan vurderes ut fra hvorvidt oppgaven kan *dekomponeres* slik at man kan skille ut en del som passer for DL-systemet, og at *kompleksiteten er håndterbar* for DL-systemet (Baird & Maruping, 2021). Dekomponeringen ble gjort i to trinn, først ved å dekomponere etter *pasientgrupper* og plukke ut hvilken pasientgruppe DL-modulen skulle brukes på, og deretter ved å dekomponere etter *strukturer* i bildeseriene og velge hvilke av strukturene som DL-modulen hadde segmentert som skulle beholdes. Man kunne tenke seg at de i stedet for å plukke ut en hel pasientgruppe, kunne plukket ut hvilke *pasienter* man skulle bruke DL-modulen på. Dette ville økt sannsynligheten for god kvalitet på all segmentering som ble gjort av DL-modulen. Imidlertid ble dette valget forklart med at segmentering nå tok så kort tid, at det var lettere å ta alle pasienter i denne gruppen gjennom DL-modulen og rette feilsegmentering etterpå, enn å gjøre vurderinger av hvilke bildeserier man trolig ville få godt segmenteringsresultat på i forkant. En annen fordel med å la DL-modulen segmentere alt først var at det segmenteres flere titalls strukturer for hver bildeserie, så selv om noe ble feilsegmentert på en eller noen strukturer, så var det tid å spare på å beholde de

andre. Når det gjelder oppgavens kompleksitet, så gikk ikke studien dypt inn på det, men det ble nevnt flere faktorer som er relevante for dette. Et par eksempler er at treningsdata stort sett kom fra de samme bildedannende enhetene[48] som ble brukt i produksjon og at det ble brukt en standard for inntegning (ESTRO Guidelines), to faktorer som begge vil redusere variasjon og dermed gjøre oppgavene mindre komplekse. På den andre siden er menneskekroppen forskjellig og derfor kilde til stor variasjon. Eksempelvis kan strukturene som skal segmenteres ha ulik størrelse fra person til person og dermed avbildes på et variabelt antall bilder, og operasjonssåret kan ligge på ulike steder i brystet. Noen slike utfordringer representerte en kompleksitet som DL-systemet klarte å håndtere, og resten tok helsepersonellet seg av. På den måten formet helsepersonellet et ansvarsområde som var innenfor DL-modulens kapabiliteter.

I Sykehus 1 var det bare ferdig utdannet helsepersonell med erfaring med manuell segmentering som fikk bruke resultatet fra DL-modulen. De ansatte fortalte at erfaring i å segmentere manuelt var en forutsetning for å både kunne kontrollere resultatet fra DL-modulen og segmentere manuelt når det var behov for det. Dette er kanskje også nødvendig for å sikre fremtidig tillit til kunstig intelligens, siden det er funnet at helsepersonell som bruker KI-systemer på medisinske bilder har mer tillit til systemet enn man ser ved annen medisinsk bruk av kunstig intelligens. Dette er forklart med at personellet kan se bildene før eller etter analysen, og har kompetanse til å analysere bildene selv for å kontrollere de KI-baserte konklusjonene (Makhlysheva et al., 2022). Tiltaket kan imidlertid også ha en konsekvens på stabilisering av delegeringssituasjonen. I henhold til Baird and Maruping (2021) og Rahwan et al. (2019) kan en stabil delegeringssituasjon både påvirke ønsket om å delegere, og hvorvidt man lykkes med å bruke produktet, og flere stabiliserende faktorer kunne jeg gjenkjenne i case 1: Helsepersonellet hadde god *kontroll på inndata*, og kjente inndata produserte *konsistente utdata* hos pasientene de hadde valgt å bruke DL-modulen på. De kunne *observere resultatet* av arbeidet til DL-modulen, og *kontrollere situasjonen* ved å endre på feil i segmenteringen ved behov. Akkurat sistnevnte ville vært veldig vanskelig for LIS-leger, som hadde mindre erfaring med segmentering. Det er derfor grunn til å tro – etter Baird and Maruping (2021) og Rahwan et al. (2019), at delegeringssituasjonen ville vært mer ustabil, og kanskje ikke fungert slik den var satt opp nå, for LIS-leger.

[48] I hvert fall halvparten av treningsdataene. Jeg kjenner ikke til detaljer knyttet til treningsdata fra det andre sykehuset som samarbeider om dette.

Figur 30 viser Sykehus 1s bruk av DL-modulen inn i en matrise med oppgavekompleksitet vertikalt og antall tilfeller teknologien brukes på, av pasienter den er trent for, horisontalt. Figuren viser at DL-modulen tok unna et stort volum av pasienter innenfor gruppen den var trent for, og at det var de standard pasientene – tilfeller med lav kompleksitet, som modulen hovedsakelig tok seg av.



Figur 30 Case 1: DL-modulen var en "arbeidsvillig assistent"

5.3.2 Case 2: Personellet får råd av en kompetent kollega

I Sykehus 2 hadde de valgt å bruke en DL-tjeneste som en likeverdig part i enkelte vanskelige situasjoner hvor de hadde behov for hjelp. Røntgenlegen var en del av et lite miljø hvor han hadde få kolleger å støtte seg til. I de fleste tilfeller var det uproblematisk og han gjorde vurderingene alene, men i enkelte tilfeller var det vanskelig å være helt sikker på at det var greit å sende pasienten hjem. Røntgenlegens oppgave var å vurdere CT-bildene av alle pasientene og avklare hvorvidt pasienten hadde sykdom av betydning. Pasienter som hadde sykdom eller vippet mot at de hadde sykdom, ble alltid sendt direkte til koronar angiografi. Friske pasienter ble sendt hjem. Var det en pasient som trolig var frisk, men røntgenlegen ikke var helt sikker på om det var greit å sende vedkommende hjem, gjorde han en vurdering av hvorvidt det var stor nok usikkerhet og kostnadsmessig riktig å bruke DL-tjenesten. I så fall sendte han CT-bildene til tjenesten. DL-tjenestens oppgave var å være beslutningsstøtte i disse situasjonene. Basert på CT-bildene laget DL-tjenesten en digital anatomisk modell av akkurat dette hjertet med koronarkar og tilhørende innsnevringer. Deretter simulerte man

blodstrømmen gjennom den anatomiske modellen, og eventuelle innsnevringer ble kalkulert, slik at det var tydelig hvor pasienten hadde eventuelle innsnevringer av funksjonell betydning. Dette ble sendt tilbake til helsepersonellet i Sykehus 2. Selv om bruken var relativt liten og svært spisset, var muligheten til å bruke DL-tjenesten likevel kjærkommen for røntgenlegen – *en kompetent kollega* som kunne gi ekspertråd og trygge røntgenlegen i egne vurderinger.

I likhet med case 1, ble også arbeidsfordelingen i dette caset påvirket av at mennesker og maskiner har ulike kapabiliteter, men det var ikke like tydelig forskjell som i case 1. Et eksempel er at menneskets syn ble nevnt som en litt begrensende faktor for helsepersonellet, men forskjellen var ikke av avgjørende betydning, og røntgenlegen fortalte også at når han hadde problemer med tolkningen ved f.eks. kalk i årene, så kunne DL-systemet også slite med det samme. En større forskjell var nok at dyp læring var brukt til å lage en anatomisk modell – en digital kopi av nettopp dette hjertet og tilhørende koronarkar, basert på tilsendte CT-bilder. Ved å gjøre simuleringer av blodstrømmen gjennom denne, kunne systemet regne seg frem til hvor det var innsnevringer av funksjonell betydning, med en langt større presisjon enn en person ville kunne gjort.

I henhold til Topol (2019) og Asan et al. (2020) er fordeler med maskinelle agenter deres evne til å prosessere store datamengder raskt og med høy presisjon. Mens det i case 1 kanskje helst var DL-systemets *hurtighet* som gav størst fordel bare presisjonen var god nok, tenker jeg at DL-systemets *presisjon* var av størst betydning i case 2, selv om en viss hastighet også var en nødvendighet. Et eksempel på når det er behov for dette er når det er tetthet i karene, og prosentvis tetthet ligger svært nær grenseverdien som røntgenlegen bruker for å skille pasienter som skal sendes hjem fra pasienter som skal videre til koronar angiografi. Det er kjent at DL-systemer kan ha høy presisjon (Asan et al., 2020; Rai, 2019), og sykehuset bruker i dette tilfellet denne kapabiliteten for å gi et svært nøyaktig svar. Å gi nøyaktige diagnoser eller svar er regnet som et lovende bruksområde for KI i helsetjenesten (Langlotz, 2019; Rajpurkar et al., 2022). Et annet eksempel er når det er litt trangt over et lengre område, og røntgenlegen er usikker på om det totalt sett er for stor reduksjon av blodstrømmen. Her er ikke problemet hvor nøyaktig man kan vurdere trangheten i ett punkt, men å regne ut total reduksjon av blodstrømmen over et lengre område, noe som er umulig for et menneske som bare ser på CT-bildene, og et eksempel på at informasjonsmengdene (punktvis tranghet over et lengre område) er for store for at mennesker klarer å nyttiggjøre seg av dem (Canals & Heukamp, 2019). DL-tjenesten har flere fordeler som den kan dra på, som verktøy for utregningen (digital anatomisk modell av hjertet), passende kapabiliteter slik at den kan finne både

mer kompliserte og subtile sammenhenger enn mennesker kan (The Lancet Digital, 2022), og bedre kapasitet til å gjøre den komplekse beregningen, som tidligere nevnt forventet fra Rai (2019) og Topol (2019).

Dette er et interessant eksempel på konfigurasjon hvor en ikke-transparent DL-tjeneste brukes til ekspertråd. Bruksområdet er noe mer komplekst enn for case 1, men selv om manglende transparens ble så vidt nevnt, så ser det ikke ut til å forårsake stor bekymring. Sykehuset har både gjennomført studie på produktet og gode erfaringer med produktet, og i henhold til Makhlysheva et al. (2022) kan evidens veie opp for manglende transparens.

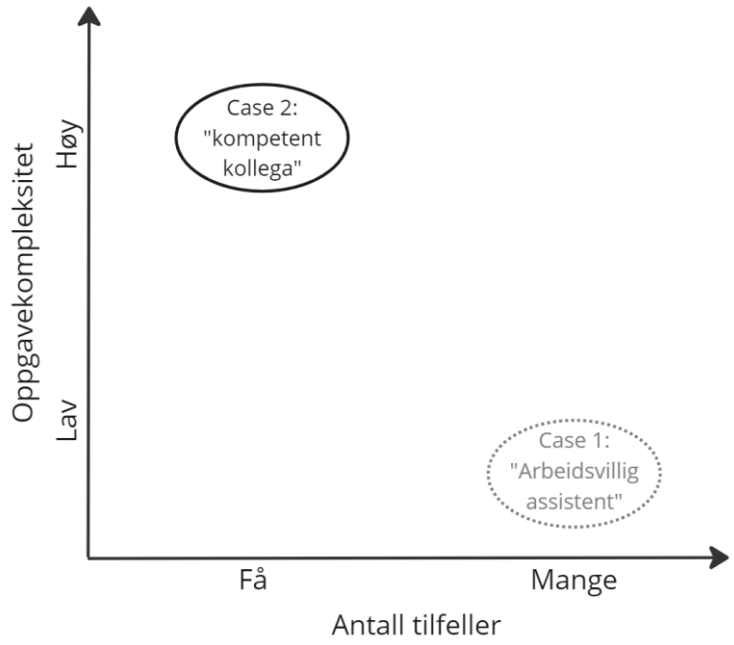
For å lage en oppgave passende for DL-tjenesten, dekomponerte helsepersonellet oppgavene som forventet fra Baird and Maruping (2021). Til forskjell fra case 1 ble dekomponeringen gjort på pasientnivå – i praksis ved å plukke ut enkeltpasienter som ellers ville gått til invasiv undersøkelse, men som man i stedet valgte å bruke DL-tjenesten på. De to oppgavene som vanligvis ble delegert til DL-tjenesten, lettgradige innsnevringer over et lengre område og innsnevring som ligger nær grenseverdien for hjemsendelse, innebærer henholdsvis komplekse beregninger og stor nøyaktighet. Selv om man i dette caset har valgt en annen type oppgaver til DL-tjenesten enn i case 1, så er det ikke uventet at systemet blir brukt til dette, siden DL-systemer kan ha høy presisjon og god kapasitet til prosessering av store datamengder (Asan et al., 2020; Topol, 2019). I tillegg til behov for nøyaktighet og komplekse beregninger, er det også andre faktorer som gjør kompleksiteten noe høyere. Pasienten kan ha høy kalsiumscore som kan gjøre det vanskelig selv for maskinsyn å tolke disse bildene, og mens man i case 1 kan forsøke å ligge så i ro som mulig for å få god kvalitet på bildene, så er det selvfølgelig ikke ønskelig at hjertet er helt i ro, selv om man roer frekvensen med medikamenter før undersøkelsen.

Når det gjelder selve delegerings*situasjonen*, så fremstod helsepersonellet trygge og med god kontroll på situasjonen. I likhet med case 1 hadde de god kontroll på inndata, forutsigbarhet i at kjent input til DL-tjenesten gir et resultat på ekspertnivå, og mulighet til å observere resultatet av DL-tjenestens arbeid ved å logge inn i en webløsning og studere 3D-modellen av hjertet og utregninger av blodstrømmen. Dette er faktorer som ifølge Baird and Maruping (2021) kan bidra til trygghet om at man kan delegere. Når det gjelder mulighet til å kontrollere delegeringssituasjonen, så var det en tydelig forskjell mellom case 1 og 2. Mens man i case 1 som regel visste hva resultatet burde være, så var ikke personellet like sikre i case 2 – det var nettopp derfor de brukte DL-tjenesten. Selv om helsepersonellet i case 2 oppfattet rådene fra DL-tjenesten som svært gode, så var det likevel mitt

inntrykk at heller ikke de stolte blindt på resultatet fra DL-systemet, men brukte det som en av flere kilder til støtte i beslutningen. Menneskers fordel ved å kunne ta med flere faktorer i beslutninger enn det KI-systemer kan, er kjent og påpekt i The Lancet Digital (2022).

Jeg fikk inntrykk av at kalibreringen av tillit til DL-tjenesten i case 2 var god, det vil si at det var god overensstemmelse mellom det intervjupersonene fortalte om DL-tjenestens kapasitet og kapabiliteter, og helsepersonellets bruk av systemet (Asan et al., 2020; Glikson & Woolley, 2020), men bruken var kanskje noe mindre enn erfaringene deres og tilliten til systemet skulle tilsi. Grunnet kostnader knyttet til hver enkelt bildeserie som ble sendt inn, måtte helsepersonellet redusere bruken til et absolutt minimum, noe som ikke er overraskende og helt på linje med Baird and Maruping (2021) som nevner økonomiske rammene som eksempel på en kognitiv ramme som spiller inn på delegeringsviljen.

Figur 31 er samme som Figur 30, men med Sykehus 2s DL-tjeneste inntegnet. Figuren viser at de to sykehusene har valgt konseptuelt forskjellige måter å bruke DL-systemene på. Matrisen viser oppgavekompleksitet vertikalt og antall tilfeller teknologien brukes på, av pasienter den er trent for, horisontalt. Case 2 brukes i "få tilfeller av pasienter den er trent for", og kun der det er "høy oppgavekompleksitet". Vi ser at ingen av konseptene plasserer seg i "lav oppgavekompleksitet" / "få tilfeller", kanskje av naturlige grunner, siden det ville vært et ulønnsomt tilfelle. Ut fra figuren ser vi at man kan få bedre hjelp av teknologien ved å gjøre den robust nok til at den på sikt kan strekkes mot øverste høyre hjørne, "høy oppgavekompleksitet" / "mange tilfeller".



Figur 31 Case 2: DL-tjenesten var en "kompetent kollega"

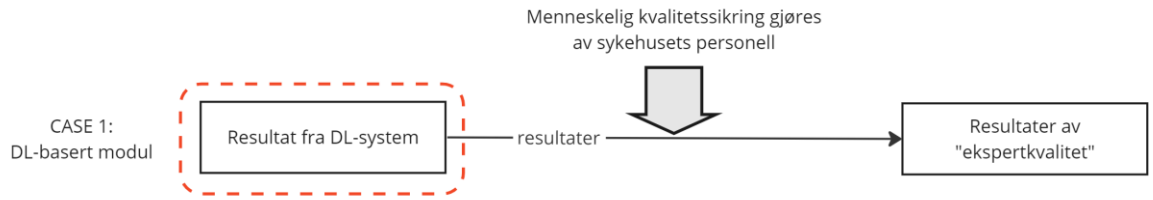
5.3.3 Pålitelige helsetjenester basert på en litt upålitelig kjerne

Det funnet som jeg mener var mest spennende er at man i begge casene har tatt i bruk beslutningsstøttesystemer for eksperter innenfor medisin – som vi vet har store krav til pålitelighet – oppå et DL-system som potensielt har flere "ukjente ukjente", usikkerheter og skjulte avhengigheter enn tradisjonelle IT-systemer (Ozkaya, 2020).

Det er også verdt å merke seg at forretningsmodellene – som virket så forskjellige på overflaten (DL-modul vs skybasert DL-tjeneste), likevel hadde en del likheter når jeg gikk inn i detaljene. Det var svært interessant å se at begge de to DL-systemene var knyttet til mennesker med *domene*kunnskap som kvalitetssikret resultatene. I case 1 gjorde sykehusets eget personale denne jobben, i case 2 gjorde leverandørens analyseteam det.

En forutsetning for delegering er at forskjellene på kapabiliteter og kunnskap er tilstrekkelig store til å rettferdiggjøre kostnadene knyttet til delegeringen (Baird & Maruping, 2021). Jeg har allerede nevnt at økonomiske faktorer begrenset Sykehus 2s bruk av DL-tjenesten, men også Sykehus 1 hadde kostnader knyttet til bruk av DL-modulen. I tillegg til en sum de betaler for å bruke produktet på et visst antall pasienter per år, var det også kostnader knyttet til kvalitetssikring av DL-modulens

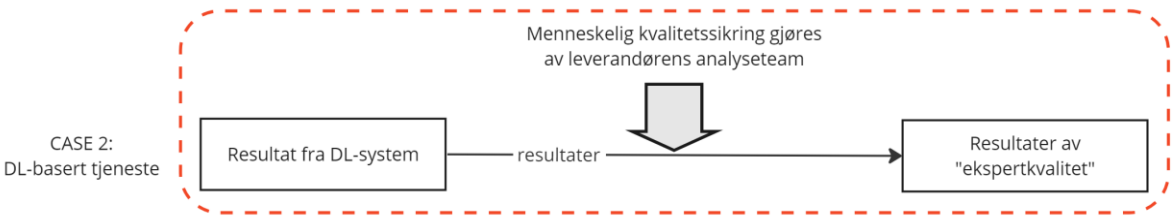
segmentering. Som beskrevet av Asatiani et al. (2021) rammet de ansatte i case 1 inn bruken av DL-systemet med klare rutiner og kvalitetssikringstiltak, for eksempel god kontroll på treningsdata, som medførte ekstraarbeid og dermed kostnader. Siden all segmentering ble kvalitetssikret av mennesker trengte ikke DL-modulen å gjøre perfekt jobb, men det er sannsynlig at modulen måtte gjøre god nok jobb til at det opplevdes som ressursbesparende å bruke den.



Figur 32 Case 1: Konseptuell skisse av hva som leveres til sykehuset

Figuren viser konseptuell skisse av hva som leveres (innenfor stiplede linjer) til sykehuset i case 1. Resultater fra DL-systemet leveres direkte til de ansatte, og den menneskelige kvalitetssikringen gjøres av sykehusets personell. Etter kvalitetssikringen og eventuell retting, kan man si at resultatene har ekspertkvalitet.

I Sykehus 2 var det leverandørens ansatte, *analyseteamet*, som gjorde det tilsvarende kvalitetssikringsarbeidet. Analyseteamet bestod av tekniske eksperter og domeneeksperter. Det er nærliggende å tro at helsepersonalet i analyseteamet og kanskje andre mennesker knyttet til ulike deler av denne tjenesten var nødvendig for å sikre at tjenesten som helhet kunne levere gode nok resultater til å gi ekspertråd. Det er ikke gjort intervjuer hos leverandøren, så dette er min tolkning av informasjon som leverandøren har lagt ut på internett.



Figur 33 Case 2: Konseptuell skisse av hva som leveres til sykehuset

Konseptuell skisse av hva som leveres (innenfor stiplede linjer) til sykehuset i case 2. Resultater fra DL-systemet kvalitetssikres av leverandørens analyseteam. Rettinger blir gjort før de returnerer resultater av ekspertkvalitet.

5.4 Implikasjoner

5.4.1 Implikasjoner for praksis

Grunn til å vurdere å ta i bruk dyp læring på medisinske bilder

Helsetjenesten vil fremover oppleve store utfordringer med tilgang på personell (Helsepersonellkommisjonen, 2023), noe man også ser konturene av i de to casene, om enn ikke i stort omfang. Casene er eksempler på at DL-systemer kan hjelpe helsepersonellet i daglig arbeid, til å frigjøre tid eller få ekspertråd. Begge casene i denne studien var innenfor radiologi. Bruk av DL-systemer innenfor radiologi er et godt område å starte utforskningen på, både fordi man har standardiserte data[49](Makhlysheva et al., 2022), stort omfang av CE-merkede produkter tilgjengelig på markedet (van Leeuwen, 2023), og forskning som viser at DL-systemer ved en del tilfeller har gitt svært gode resultater på medisinske områder (Asatiani et al., 2021; Esteva et al., 2017; Zaharchuk & Davidzon, 2021), spesielt innenfor spesialiteter som lener seg tungt på tolkning av bilder (Zaharchuk & Davidzon, 2021). Gitt de økte utfordringene knyttet til personellsituasjonen i helsetjenesten (Helsepersonellkommisjonen, 2023) er det derfor allerede nå grunn til å teste og eventuelt ta i bruk DL-systemer, i hvert fall innen medisinske fagdisipliner som bruker digitale medisinske bilder.

Lage "fasit" for medisinske bilder er tidkrevende og komplisert

Fra teorien vet vi at det kan være vanskelig å definere en helt korrekt "fasit" for treningsdata – *ground truth* (Asan et al., 2020), noe som gav utslag i case 1 hvor de erfarte at det tok lang tid å lage segmenterte treningsdata blant annet fordi det er noe variasjon i hvordan leger segmenterer. Selv om Sykehus 2 ikke trente modellen selv, kom de heller ikke helt unna problemstillingen, siden de måtte etablere en "fasit" for å ettergå resultatene fra DL-modulen. Hvordan trenings- eller testdata skal lages og hvilke mengder data man trenger må derfor tas med i betraktningen når man beslutter å ta i bruk kunstig intelligens, og kan være en faktor av betydning inn i beslutning om hvorvidt man skal anskaffe ferdig trent DL-system eller trene modellen selv.

[49] Digital Imaging and Communications in Medicine (DICOM®), https://www.dicomstandard.org/about

Behov for dobbel kompetanse

I Sykehus 1 hadde flere intervjupersoner et aktivt forhold til å sikre fremtidig kompetanse innenfor segmentering, og leger i spesialisering[50] (LIS) måtte derfor lære å segmentere manuelt før de fikk bruke DL-modulen. Det er kjent motsetningsforhold at DL-systemer på den ene siden kan forbedre beslutningskvaliteten til et team, men på den andre siden bidra til redusert evne til kritisk tenkning (Seeber et al., 2020), noe som dette tiltaket i Sykehus 1 kan bøte på. I tillegg til medisinsk fagkompetanse må de ansatte ha tilstrekkelig kompetanse om dyp læring til å skjønne hvilke hendelser som kan oppstå slik at virksomheten kan sette i verk risikoreduserende tiltak knyttet til bruken av systemet. Dette betyr at det er behov for dobbel kompetanse. I en artikkel om utvikling av nye legemidler var ansatte med dobbel kompetanse (medisinsk + KI) avgjørende for å bli gode på innovasjon ved hjelp av DL-baserte verktøy. Dette ble forklart med behovet for kontinuerlig syntese av kunnskap fra både DL- og medisinekspertise (Lou & Wu, 2021). Det kan tenkes at dette er gjeldende for andre arbeidsoppgaver innenfor medisin og at dobbel kompetanse kan være attraktivt å se etter ved ansettelsesprosesser.

Må venne seg til å lage stabile, sikkerhetskritiske tjenester basert på en upålitelig kjerne

Tradisjonell kvalitetssikring kan være både utilstrekkelig og uegnet for DL-systemer innen medisin av flere grunner. Medisin er et komplekst, sikkerhetskritisk domene (Habli et al., 2020), og DL-systemer er knyttet til utfordringer som må håndteres, inkludert at resultatene er vanskelige å granske, forstå og forklare på grunn av store treningsdata, tette bånd til dataene og skjulte prosesseringslag (Glikson & Woolley, 2020; Lyytinen et al., 2021; Teodorescu et al., 2021). I denne studien var det få tilfeller av denne typen utfordringer, bortsett fra sjeldne tilfeller i case 2 der DL-tjenesten ga uventede svar uten klar årsak. Vi så at sykehuset i case 1 hadde en rutinemessig jobb i å kvalitetssikre resultatene fra DL-modulen. At kvalitetssikring ikke ble problematisert i særlig grad i case 2 kan ha mange årsaker, og er også vanskelig å tolke siden jeg ikke har full innsikt i hva som skjer av kvalitetssikring på leverandørsiden. Det kan på den ene siden tyde på at DL-systemet i kjernen av tjenesten var robust og gav stabilt gode svar, men det kan like gjerne bety at leverandøren har en stor kvalitetssikringsjobb før de sender ut resultater.

[50] Leger som læres opp

Det er kjent at DL-systemene kan være usikre, ha skjulte avhengigheter og mange "ukjente ukjente" (Ozkaya, 2020), og de kan være lette å forvirre (Castelvecchi, 2016). I case 1 ble det nevnt at DL-modulen kunne *"bli litt forvirret og lage et feil resultat"* (f) hvis den ble brukt på data utenfor treningssettet. Sikker verifisering er en utfordring, og velkjente arkitekturprinsipper for å kontrollere avhengigheter er vanskelige å anvende på disse systemene (Ozkaya, 2020; Teodorescu et al., 2021). Helsetjenesten har begrenset erfaring med DL-systemer (Rajpurkar et al., 2022) og det vil ta betydelig tid og ressurser å sikre trygg bruk i praksis, som for eksempel i case 1 hvor de brukte mye tid på å lage treningsdata og i case 2 hvor de gjennomførte invasiv studie for alle pasienter i en periode. Ved eventuell fremtidig bruk av DL-systemer som lærer i produksjon vil kvalitetssikringen kunne bli svært utfordrende. Tradisjonelle metoder er ikke tilstrekkelige til å håndtere disse utfordringene, selv med omfattende testing. Teknikker for å ramme inn bruken, øke forklarbarhet og forståbarhet kan være utilstrekkelige når systemet er trent på store treningsdata som mennesker ikke klarer å få oversikt over. Man må derfor tenke nytt rundt kvalitetssikring for å sikre pålitelige helsetjenester basert på upålitelig teknologi.

5.4.2 Implikasjoner for videre forskning

Det er lite forskning på hvordan mennesker og KI-systemer spiller sammen (Grønsund & Aanestad, 2020), delegering av ansvar mellom KI-systemer og spesialister i helsetjenesten (Baird & Maruping, 2021) og forskning på kunstig intelligens i virkelige miljøer (Glikson & Woolley, 2020) blant annet i spesialisthelsetjenesten (Asatiani et al., 2021). Denne studien utgjør et lite bidrag, men mange temaer har jeg bare så vidt skrapet i overflaten på. Her presenterer jeg to mulige områder for videre forskning.

Nye samarbeidsformer og tjenestemodeller

Kunstig intelligens vil trolig brukes på ulike måter i sykehusene, som vist i de to casene i denne studien, og kan studeres i ulike perspektiver. Det er helt nødvendig å ha tilgang til treningsdata for utvikling av DL-systemer (Ozkaya, 2020), men det kan være utfordrende å få tak i gode og nok data fra helsetjenesten for leverandøren (Makhlysheva et al., 2022). Det er derfor interessant å følge med på i hvilken grad DL-systemenes uløselige forhold til data kan bli en driver for nye samarbeidsformer og forretningsmodeller for medisinsk bruk av DL-systemer.

Forskning på tjenestemodeller for kunstig-intelligens-baserte systemer i et sosioteknisk perspektiv er interessant. I case 2 ser vi kanskje konturene av en variant av DL-baserte tjenestemodeller. Det interessante i case 2 er at *domene*eksperter, "trent personell", fra leverandøren ble nevnt som en del av analyseteamet i tjenesten som ble levert. I begge casene var det behov for å holde resultatene fra DL-modulen under oppsyn. Det er interessant å få mer klarhet i hvorvidt DL-systemer – per nå eller av natur – er for uforutsigbare og upålitelige til at de alene kan levere ekspertråd når det gjelder liv og helse. I så fall kan vi få behov for en sosioteknisk tilleggsdimensjon til den vanlige "as-a-service"-modellen, "Sosiotechnical AI System as a Service", hvor mennesker har som rolle å korrigere resultatene fra DL-systemet før de leveres ut av tjenesten. I motsatt fall kan det tenkes at dette blir unødvendig dersom teknologien blir mer robust og systemene etter hvert trenes med så store mengder treningsdata at systemene oppleves like pålitelige som dagens tradisjonelle IT-systemer.

Bruke upålitelig teknologi for å lage pålitelige helsetjenester

Pålitelighet er et tema som bør utforskes videre, og da med fokus på spenningen mellom dyp lærings litt upålitelige natur på den ene siden, og livsviktig pasientsikkerhet på den andre siden. Dette bør utforskes i flere perspektiver, som psykologisk trygghet, og tillit mellom pasient og helsetjeneste. Hva som bidrar til trygghet for de som bruker DL-systemer er interessant å utforske både i sosioteknisk og kvalitetssikringsperspektiv. Grunnet DL-systemenes upålitelige natur kan det være grunn til å studere bruk av DL-systemer i lys av teori som har et element av uforutsigbarhet i seg, som beredskap, innovasjon, disrupsjon og risikostyring – sistnevnte er også naturlig fordi det er tilnærmingen som brukes i AI act[51] som trolig vil sette rammer for bruken av KI i Norge.

Det kunne vært interessant å undersøke i hvilken grad tilliten man har til et DL-produkt er relatert til kundens involvering i utviklingsprosessen, for eksempel ved å undersøke andre kunders tillit til samme produktet. Selv om man i case 1 i denne studien har godt kalibrert tillit til DL-modulen, kan man tenke seg at andre kunder som ikke har vært med å trene modellen har lavere tillit til DL-modulen. Dette vil være i forlengelsen av eksempelvis Baird & Maruping (2021) som hevder at åpenhet rundt den maskinelle agentens innebygde preferanser er viktig for et godt forhold mellom den menneskelige og den maskinelle agenten.

[51] https://artificialintelligenceact.eu

6 Konklusjon

I denne studien studerer jeg erfaringer fra bruk av kunstig intelligens i spesialisthelsetjenesten ved å svare på forskningsspørsmålet *"hvordan skjer sosioteknisk rekonfigurering ved innføring av KI-system[52] i sykehus"*, ved å belyse motivasjon for innføringen av det KI-baserte systemet, viktige endringer og særtrekk ved ny konfigurasjon.

En sentral motivasjonsfaktor for innføringen var ressursutfordringer og forventning om at DL-systemene kunne bøte på disse. Kvalitetssikring og hensyn til pasienten stilte også en stor rolle.

DL-systemets uløselige forhold til sine treningsdata la føringer for endringer og tiltak for å ta DL-systemet i daglig bruk, og jeg stiller spørsmål ved hvorvidt kjennskap til treningsdata er en fordel eller forutsetning for trygg bruk i helsetjenesten slik DL-teknologien er i dag.

Helsepersonellet hadde siste ordet i begge konfigurasjonene, men DL-systemene ble brukt konseptuelt forskjellig. I case 1 brukte personellet DL-modulen som *en arbeidsvillig assistent*, og i case 2 brukte de DL-tjenesten som *en kompetent kollega*. Jeg påpeker spenning mellom DL-systemets upålitelige natur og bruk på et så sikkerhetskritisk domene som medisin, og at helsetjenesten i fremtiden trolig på rutinemessig basis må håndtere det å levere pålitelige helsetjenester basert på upålitelig teknologi.

Studien er et bidrag til litteraturen, siden det hittil har vært lite forskning på erfaringer fra bruk av DL-systemer i ekte, komplekse situasjoner, både generelt (Glikson & Woolley, 2020) og i spesialisthelsetjenesten (Asatiani et al., 2021). Studien har også trolig implikasjoner for praksis, siden den indikerer at sykehusene bør fortsette å ta i bruk DL-baserte systemer, til tross for at det vil kunne medføre en rekke endringer, som å etablere ny praksis for kvalitetssikring.

[52] KI-system er et kunstig-intelligens-basert system, for eksempel basert på dyp læring (DL) som de to casene i denne studien

Referanser

Arnesen, H., Holck, P., & Hisdal, J. (2022). Hjertet. *Store medisinske leksikon på snl.no*. Retrieved 11 07, from https://sml.snl.no/hjertet

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of medical Internet research*, *22*(6), e15154-e15154. https://doi.org/10.2196/15154

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2021). Sociotechnical envelopment of artificial intelligence: an approach to organizational deployment of inscrutable artificial intelligence systems. *Journal of the Association for Information Systems*, *22*(2), 325-352. https://doi.org/10.17705/1jais.00664

Baird, A., & Maruping, L. M. (2021). The next generation of research on is use: A theoretical framework of delegation to and from agentic is artifacts. *MIS quarterly*, *45*(1), 315-341. https://doi.org/10.25300/MISQ/2021/15882

Bechky BA, O. M. S. (2015). Leveraging comparative field data for theory generation In (1 ed.). Routledge. https://doi.org/https://doi.org/10.4324/9781315849072

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing Artificial Intelligence. *MIS quarterly*, *45*, 1433-1450. https://doi.org/10.25300/MISQ/2021/16274

Bertuzzi, L. (2023, 07.03.2023). EU lawmakers set to settle on OECD definition for Artificial Intelligence. *EURACTIV*. https://www.euractiv.com/section/artificial-intelligence/news/eu-lawmakers-set-to-settle-on-oecd-definition-for-artificial-intelligence

Brekke, M. (2018). Radiolog. *Store medisinske leksikon på snl.no*. Retrieved 12 17, from https://sml.snl.no/radiolog

Brekke, M., & Borthne, A. (2022a). Radiograf. *Store medisinske leksikon på snl.no*. Retrieved 02 15, from https://sml.snl.no/radiograf

Brekke, M., & Borthne, A. (2022b, 14.01.2022). *Radiologi*. Store medisinske leksikon. Retrieved 30.05.2023 from https://sml.snl.no/radiologi

Brekke, M., Kolbenstvedt, A., & Borthne, A. (2022). CT. *Store medisinske leksikon på snl.no*. Retrieved 04 26, from https://sml.snl.no/CT

Brinkmann, S. (2012). *Qualitative inquiry in everyday life*. SAGE.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of behavioral decision making*, *33*(2), 220-239. https://doi.org/10.1002/bdm.2155

Canals, J., & Heukamp, F. (2019). How Can Human-Computer "Superminds" Develop Business Strategies? In (pp. 165-183). Switzerland: Springer International Publishing AG. https://doi.org/10.1007/978-3-030-20680-2_9

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, *538*(7623), 20-23. https://doi.org/10.1038/538020a

Coiera, E. (2004). Four rules for the reinvention of health care. *BMJ*, *328*(7449), 1197-1199. https://doi.org/10.1136/bmj.328.7449.1197

Coiera, E. (2020). The cognitive health system. *Lancet*, *395*(10222), 463-466. https://doi.org/10.1016/S0140-6736(19)32987-3

Curry, D. (2023). ChatGPT Revenue and Usage Statistics [29.05.2023]. Retrieved 05.05.2023, from https://www.businessofapps.com/data/chatgpt-statistics

Engebretsen, E. B., Hilde;. (2023, 22. mai 2022). *Kunnskapsbasert medisin i Store medisinske leksikon på snl.no*. Retrieved 9. april 2023 from https://sml.snl.no/kunnskapsbasert_medisin

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115-118. https://doi.org/10.1038/nature21056

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *The Academy of Management annals*, *14*(2), 627-660. https://doi.org/10.5465/annals.2018.0057

Golden-Biddle, K., & Locke, K. (2007). *Composing qualitative research* (2nd ed. ed.). Sage.

Grønsund, T., & Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. *The Journal of Strategic Information Systems*, *29*(2), 101614. https://doi.org/10.1016/j.jsis.2020.101614

Habli, I., Lawton, T., & Porter, Z. (2020). Artificial intelligence in health care: Accountability and safety. *Bull World Health Organ*, *98*(4), 251-256. https://doi.org/10.2471/BLT.19.237487

Helsedirektoratet. (2022, 10.02.2023). *Myndigheters ansvar for ulike regelverk*. Helsedirektoratet. Retrieved 29.05.2023 from https://www.helsedirektoratet.no/tema/kunstig-intelligens/regelverk/myndigheters-ansvar-for-ulike-regelverk

Helsepersonellkommisjonen. (2023). *Tid for handling - Personellet i en bærekraftig helse- og omsorgstjeneste* https://www.regjeringen.no/contentassets/337fef958f2148bebd326f0749a1213d/no/pdfs/nou202320230004000dddpdfs.pdf

Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the Quality of Explanations: The System Causability Scale (SCS). *KI - Künstliche Intelligenz*, *34*(2), 193-198. https://doi.org/10.1007/s13218-020-00636-z

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov*, *9*(4), e1312-n/a. https://doi.org/10.1002/widm.1312

Huo, W., Zheng, G., Yan, J., Sun, L., & Han, L. (2022). Interacting with medical artificial intelligence: Integrating self-responsibility attribution, human–computer trust, and personality. *Computers in human behavior*, *132*, 107253. https://doi.org/10.1016/j.chb.2022.107253

Johnson, J. L., Adkins, D., & Chauvin, S. (2020). A review of the quality indicators of rigor in qualitative research. *Am J Pharm Educ*, *84*(1), 138-146. https://doi.org/10.5688/ajpe7120

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15-25. https://doi.org/10.1016/j.bushor.2018.08.004

Kitamura, F. C. (2023). ChatGPT Is Shaping the Future of Medical Writing But Still Requires Human Judgment. *Radiology*, *307*(2), e230171-e230171. https://doi.org/10.1148/radiol.230171

Klein, H. K., & Myers, M. D. (1999). A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems. *MIS quarterly*, *23*(1), 67-93. https://doi.org/10.2307/249410

Kleppe, A., Skrede, O.-J., De Raedt, S., Hveem, T. S., Askautrud, H. A., Jacobsen, J. E., Church, D. N., Nesbakken, A., Shepherd, N. A., Novelli, M., Kerr, R., Liestøl, K., Kerr, D. J., & Danielsen, H. E. (2022). A clinical decision support system optimising adjuvant chemotherapy for colorectal cancers by integrating deep learning and pathological staging markers: a development and validation study. *The lancet oncology*, *23*(9), 1221-1232. https://doi.org/10.1016/S1470-2045(22)00391-6

Kleppe, A., Skrede, O.-J., De Raedt, S., Liestøl, K., Kerr, D. J., & Danielsen, H. E. (2021). Designing deep learning studies in cancer diagnostics.

Kvale, S., Brinkmann, S., Anderssen, T. M., & Rygge, J. (2015). *Det kvalitative forskningsintervju* (3. utg. ed.). Gyldendal akademisk.

Langlotz, C. P. (2019). Will Artificial Intelligence Replace Radiologists? *Radiol Artif Intell*, *1*(3), e190058-e190058. https://doi.org/10.1148/ryai.2019190058

Lee, Y. S., & Siemsen, E. (2017). Task Decomposition and Newsvendor Decision Making [Article]. *Management Science*, *63*, 3226+. https://link-gale-com.ezproxy.uio.no/apps/doc/A515125213/AONE?u=oslo&sid=bookmark-AONE&xid=3adf7b36

Leonardi, P. M. (2011). When Flexible Routines Meet Flexible Technologies: Affordance, Constraint, and the Imbrication of Human and Material Agencies. *MIS quarterly*, *35*(1), 147-167. https://doi.org/10.2307/23043493

LIS. (2022, 03/14/2022). https://spesialisthelsetjenesten.no/lis

Lou, B., & Wu, L. (2021). AI on drugs: Can artificial intelligence acceleratedrug development? Evidence from a large-scale examination of bio-pharma firms. *MIS quarterly*, *45*(3), 1451-1482. https://doi.org/10.25300/MISQ/2021/16565

Lyytinen, K., Nickerson, J. V., & King, J. L. (2021). Metahuman systems = humans + machines that learn. *Journal of Information Technology*, *36*(4), 427-445. https://doi.org/10.1177/0268396220915917

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: an integrative review. *Theoretical issues in ergonomics science*, *8*(4), 277-301. https://doi.org/10.1080/14639220500337708

Makhlysheva, A., Marco-Ruiz, L., Olsen Svenning, T., Dinh Ngo, P., Tejedor Hernandez, M., Nordsletta, A., & Tayefi, M. (2022). *Implementation of artificial intelligence in Norwegian healthcare: The road to broad adoption*. https://ehealthresearch.no/files/documents/Rapporter/NSE-rapport_2022-01_Implementation-of-AI.pdf

Mazmanian, M., Cohn, M., & Dourish, P. (2014). Dynamic Reconfiguration in Planetary Exploration

A Sociomaterial Ethnography. *MIS quarterly*, *38*(3), 831-848. https://www-jstor-org.ezproxy.uio.no/stable/26635000

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, *27*(4), 12-12.

Mees-Buss, J., Welch, C., & Piekkari, R. (2022). From Templates to Heuristics: How and Why to Move Beyond the Gioia Methodology. *Organizational research methods*, *25*(2), 405-429. https://doi.org/10.1177/1094428120967716

Mintz, Y., & Brodie, R. (2019). Introduction to artificial intelligence in medicine. *Minim Invasive Ther Allied Technol*, *28*(2), 73-81. https://doi.org/10.1080/13645706.2019.1575882

Morgan, D. L., & Nica, A. (2020). Iterative Thematic Inquiry: A New Method for Analyzing Qualitative Data. *International Journal of Qualitative Methods*, *19*, 1609406920955118. https://doi.org/10.1177/1609406920955118

Nasjonal helse- og sykehusplan 2020-2023. ((2019).

Nasjonal strategi for kunstig intelligens. (2020). https://www.regjeringen.no/no/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594

Nedelkoska, L., & Quintini, G. (2018). Automation, skills use and training. https://doi.org/doi:https://doi.org/10.1787/2e2f4eea-en

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med*, *375*(13), 1216-1219. https://doi.org/10.1056/NEJMp1606181

Okhuysen, G. A., & Bechky, B. A. (2009). 10 coordination in organizations: An integrative perspective. *Academy of Management Annals*, *3*(1), 463-502.

Ozkaya, I. (2020). What Is Really Different in Engineering AI-Enabled Systems? *IEEE Software*, *37*(4), 3-6. https://doi.org/10.1109/MS.2020.2993662

Pan, S. L., & Tan, B. (2011). Demystifying case research: A structured–pragmatic–situational (SPS) approach to conducting case studies. *Information and Organization*, *21*(3), 161-176. https://doi.org/10.1016/j.infoandorg.2011.07.001

Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health systems. *J Glob Health*, *8*(2), 020303-020303. https://doi.org/10.7189/jogh.08.020303

Payne, G. (2007). Social Divisions, Social Mobilities and Social Research: Methodological Issues after 40 Years. *Sociology (Oxford)*, *41*(5), 901-915. https://doi.org/10.1177/0038038507080444

Pratt, M. G., Kaplan, S., & Whittington, R. (2020). Editorial Essay: The Tumult over Transparency: Decoupling Transparency from Replication in Establishing Trustworthy Qualitative Research. *Administrative science quarterly*, *65*(1), 1-19. https://doi.org/10.1177/0001839219887663

Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and information technology*, *20*(1), 5-14. https://doi.org/10.1007/s10676-017-9430-8

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. S., Roberts, M. E., Shariff, A., Tenenbaum, J. B., & Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477-486. https://doi.org/10.1038/s41586-019-1138-y

Rai, A., Constantinides, P., & Sarker, S. (2019). Next Generation Digital Platforms: Toward Human-AI Hybrids. *MIS quarterly*, *43*(1), iii-ix. https://misq.org/misq/downloads/

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, *28*(1), 31-38. https://doi.org/10.1038/s41591-021-01614-0

Rangachari, P., & L. Woods, J. (2020). Preserving Organizational Resilience, Patient Safety, and Staff Retention during COVID-19 Requires a Holistic Consideration of the Psychological Safety of Healthcare Workers. *International Journal of Environmental Research and Public Health*, *17*(12), 4267. https://www.mdpi.com/1660-4601/17/12/4267

Recht, M., & Bryan, R. N. (2017). Artificial Intelligence: Threat or Boon to Radiologists? *J Am Coll Radiol*, *14*(11), 1476-1480. https://doi.org/10.1016/j.jacr.2017.07.007

Regelverk og søknader. (2022). https://www.helsedirektoratet.no/tema/kunstig-intelligens/regelverk/regelverk-og-soknader

Regjeringen. (2014). Regjeringen. *Grunnstrukturen i helsetjenesten*. Retrieved 30/10/2022, from https://www.regjeringen.no/no/tema/helse-og-omsorg/sykehus/vurderes/grunnstrukturen-i-helsetjenesten/id227440/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-Agnostic Interpretability of Machine Learning. https://doi.org/10.48550/arxiv.1606.05386

Robbins, S. (2020). AI and the path to envelopment: Knowledge as a first step towards the responsible regulation and use of AI-powered machines. *AI & SOCIETY*, *35*(2), 391-400.

Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. *Autonomous agents and multi-agent systems*, *33*(6), 673-705. https://doi.org/10.1007/s10458-019-09408-y

Salwei, M. E., & Carayon, P. (2022). A Sociotechnical Systems Framework for the Application of Artificial Intelligence in Health Care Delivery. *Journal of Cognitive Engineering and Decision Making*, *16*(4), 194-206. https://doi.org/10.1177/15553434221097357

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & management*, *57*(2), 103174. https://doi.org/10.1016/j.im.2019.103174

Shimizu, H., & Nakayama, K. I. (2020). Artificial intelligence in oncology. *Cancer Sci*, *111*(5), 1452-1460. https://doi.org/10.1111/cas.14377

Sloane, E. B., & J. Silva, R. (2020). Chapter 83 - Artificial intelligence in medical devices and clinical decision support systems. In E. Iadanza (Ed.), *Clinical Engineering Handbook (Second Edition)* (pp. 556-568). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-813467-2.00084-5

Sperre Saunes, I., Karanikolos, M., Sagan, A., & Organization, W. H. (2020). Norway: health system review.

Stokes, F., & Palmer, A. (2020). Artificial Intelligence and Robotics in Nursing: Ethics of Caring as a Guide to Dividing Tasks Between AI and Humans. *Nursing philosophy*, *21*(4), e12306-n/a. https://doi.org/10.1111/nup.12306

Suchman, L. A. (2007). *Human-machine reconfigurations: Plans and situated actions*. Cambridge university press.

Teodorescu, M. H. M., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of fairness in automation require a deeper understanding of human–ml augmentation. *MIS quarterly*, *45*(3), 1483-1499. https://doi.org/10.25300/MISQ/2021/16535

The Lancet Digital, H. (2022). Holding artificial intelligence to account. *Lancet Digit Health*, *4*(5), e290-e290. https://doi.org/10.1016/S2589-7500(22)00068-1

The Radboud university medical center. (2023). *AI for Radiology*. The Radboud university medical center. Retrieved 18.06.2023 from https://grand-challenge.org/aiforradiology

Ting, D. S. W., Cheung, C. Y.-L., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I. Y., Lee, S. Y., Wong, E. Y. M., Sabanayagam, C., Baskaran, M., Ibrahim, F., Tan, N. C., Finkelstein, E. A., Lamoureux, E. L., Wong, I. Y., Bressler, N. M., Sivaprasad, S., Varma, R., Jonas, J. B., He, M. G., Cheng, C.-Y., Cheung, G. C. M., Aung, T., Hsu, W., Lee, M. L., & Wong, T. Y. (2017). Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA*, *318*(22), 2211-2223. https://doi.org/10.1001/jama.2017.18152

Tjora, A. H. (2021). *Kvalitative forskningsmetoder i praksis* (4. utgave. ed.). Gyldendal.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*, *25*(1), 44-56. https://doi.org/10.1038/s41591-018-0300-7

Ullman, D., & Malle, B. (2017, 2017). Human-Robot Trust: Just a Button Press Away.*HRI '17* ACM/IEEE International Conference on Human-Robot Interaction,

van Leeuwen, K. G. (2023). *AI for radiology - an implementation guide*. Radboud university medical center. Retrieved 17.05 from https://grand-challenge.org/aiforradiology/

van Leeuwen, K. G., Schalekamp, S., Rutten, M. J. C. M., van Ginneken, B., & de Rooij, M. (2021). Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *European Radiology*, *31*(6), 3797-3804. https://doi.org/10.1007/s00330-021-07892-z

Véras, D., Prudêncio, R., Ferraz, C., Bispo, A., & Prota, T. (2015, 4-7 Nov. 2015). Context-Aware Techniques for Cross-Domain Recommender Systems. 2015 Brazilian Conference on Intelligent Systems (BRACIS),

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future generation computer systems*, *135*, 364-381. https://doi.org/10.1016/j.future.2022.05.014

Yu, A. C., Mohajer, B., & Eng, J. (2022). External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. *Radiol Artif Intell*, *4*(3), e210064-e210064. https://doi.org/10.1148/ryai.210064

Zaharchuk, G., & Davidzon, G. (2021). Artificial Intelligence for Optimization and Interpretation of PET/CT and PET/MR Images. *Semin Nucl Med*, *51*(2), 134-142. https://doi.org/10.1053/j.semnuclmed.2020.10.001

Vedlegg 1: Intervjuguide

1. **FAKTA OM PRODUKTET**
    - Hvilket DL-basert produkt bruker dere?
    - Er produktet kjøpt inn eller egenutviklet?
    - Hvor er produktet installert (på eller utenfor sykehuset)
    - Dersom innkjøpt:
        - Hvilken versjon av produktet bruker dere?
    - Dersom egenutviklet:
        - Vet du hvem som har utviklet det?

2. **BRUK AV KUNSTIG INTELLIGENS**
    - Hva var motivasjonen for / målet med å ta i bruk dette produktet?
    - Har dere, etter din mening, lykkes med å nå målet med å ta i bruk produktet?
    - Kva brukes kunstig intelligens til i dette produktet?
        - Er det flere ting KI brukes til i dette produktet?
        - Bruker dere all DL-basert funksjonalitet som tilbys i dette produktet?
    - Hvordan gir produktet verdi for leverandøren?
    - Hva er betalingsmodellen (engangssum, betaling pr. pasient,...)
    - Hvordan skal produktet gi verdi til helseforetaket? Hva skal det hjelpe helseforetaket med?
    - Bruker dere produktet slik det er ment fra leverandørens side? Dersom ikke: Fortell om hva som er forskjellen:
        - Hvordan vil leverandøren at produktet skal brukes?
        - Hvordan bruker helseforetaket det?
        - Hvorfor har helseforetaket valgt å benytte produktet på en annen måte?

3. **HVORDAN ER DET TEKNISK INTEGRERT?**
    - Hvordan er produktet som er basert på KI integrert i øvrig teknisk utstyr?
    - Hvordan synes du at integrasjonen fungerer, rent teknisk?
    - I hvilken grad kan helseforetaket integrere egne DL-modeller i med produktet?

- I hvilken grad kan ulike leverandørers DL-modeller integreres / brukes sammen med produktet?

4. **HVORDAN ER DET INTEGRERT I ARBEIDSFLYT/BESLUTNINGSPROSESS?**
    - Kan du fortelle hovedtrinnene i arbeidsprosessen du gjennomfører?
    - Ha er de viktigste programmene du bruker i din arbeidshverdag?
    - Hvordan er det DL-baserte produktet integrert i denne arbeidsprosessen?
    - Har du og produktet likt datagrunnlag for vurderingene/beslutningene som gjøres?

    Dersom ikke;
        - Hva har helsepersonellet, som produktet ikke har?
        - Hva har produktet, som helsepersonellet ikke har?
    - I hvilken grad, og hvordan, er arbeidsdagen din endret ved bruk av produktet?
    - I hvilken grad påvirkes du og dine beslutninger av resultatet fra produktet?
        - Hvordan?
        - Har dette endret seg over tid?
    - I hvilken grad kan du påvirke/påvirker du resultatet fra produktet?
        - Hvordan?
        - Har dette endret seg over tid?
    - I hvilken grad hender det at produktet og din vurdering er forskjellig?
        - Fortell hva som skjer da
    - Dersom forskjell:
        - Hva tror du er grunnen til at produktets og din vurdering er forskjellig?
        - Har du fått noen reaksjoner, positive eller negative, på dette?

5. **ERFARINGER**
    - Hvordan bruker du resultatene fra dette produktet/denne modulen?
    - Hva gjør du om du er usikker på rådet du får fra produktet/modulen?

6. **ENDRINGER**
    - I hvilken grad har du, kollegene dine eller andre ved helseforetaket gjort endringer for å tilpasse dere produktet?
        - Hvorfor

- Hvilken type endringer
- Hva er erfaringene dine med endringene?

7. **EVENTUELT**

- Er det annet som du tenker vi bør vite om, relatert til temaene vi har intervjuet deg om?

Vedlegg 2: Opplysninger om intervjuer

Opplysninger	**Case 1**					**Case 2**			
Tid for datainnsamling	27/5 - 9/6 2022 Verifikasjon: 22/11 Demo: 23/11					20/6 - 27/6 Verifikasjon/intervju: 2/11 - 8/11			
Type samtale	**Intervju og oppfølgingsspørsmål**	**Intervju**	**Intervju**	**Intervju**	**Intervju**	**Intervju**	**Intervju**	**Korte intervjuer / gjennomgang av figurene**	
Roller intervjuet (og forkortelse brukt ved sitater)	**Seksjons-leder (sl)**	**Stråle-terapeut (st)**	**Fysiker (f)**	**LIS-lege (l)**	**Overlege/ onkolog (o)**	**Røntgenlege (rl)**	**Intervensjons-kardiolog (ik)**	**Radiograf (r)**	**IT-radiograf (ir)**
Hvem som svarte på hvert spørsmål:									
Hvilket DL-basert produkt bruker dere?	-	x	-	-	-	x	-	-	-
Er produktet kjøpt inn eller egenutviklet?	x	x	-	-	-	x	-	-	-
Hvor er produktet installert (på eller utenfor sykehuset)	x	x	-	-	-	x	-	-	-
Dersom innkjøpt: Hvilken versjon av produktet bruker dere?	x	x	-	-	-	x	-	-	-
Dersom egenutviklet: Vet du hvem som har utviklet det?	-	-	x	-	-	-	-	-	-
Hva var motivasjonen for / målet med å ta i bruk dette produktet?	x	x	x	x	-	x	x	-	-
Har dere, etter din mening, lykkes med å nå målet med å ta i bruk produktet?	x	x	x	x	-	x	x	-	x

Hva brukes kunstig intelligens til i dette produktet?	x	x	x	x	-	x	x	-	-
Hvordan gir produktet verdi for leverandøren?	x	x	x	-	-	x	x	-	-
Hva er betalingsmodellen (engangssum, betaling pr. pasient,...)	x	x	x	-	-	x	-	-	-
Hvordan skal produktet gi verdi til helseforetaket? Hva skal det hjelpe helseforetaket med?	x	x	x	x	-	x	x	-	-
Bruker dere produktet slik det er ment fra leverandørens side? Dersom ikke: Fortell om hva som er forskjellen	x	x	x	x	-	x	x	-	-
Hvordan er produktet som er basert på KI integrert i øvrig teknisk utstyr?	x	x	x	x	-	x	-	x	-
Hvordan synes du at integrasjonen fungerer, rent teknisk?	x	x	x	-	-	x		x	-
I hvilken grad kan helseforetaket integrere egne DL-modeller i med produktet, og i hvilken grad kan ulike leverandørers DL-modeller integreres / brukes sammen med produktet?	x	x	x	-	-	x		-	-
Kan du fortelle hovedtrinnene i arbeidsprosessen du gjennomfører?	x	x	x	x	x	x	x	(x)	x
Hva er de viktigste programmene du bruker i din arbeidshverdag?	x	x	x	-	x	x	x	-	x

Hvordan er det DL-baserte produktet integrert i denne arbeidsprosessen?	x	x	x	-	x	x	x	(x)	(x)
Har du og produktet likt datagrunnlag for vurderingene/beslutningene som gjøres? (+ underspm.)	x	x	x	x	x	x	x	-	-
I hvilken grad, og hvordan, er arbeidsdagen din endret ved bruk av produktet?	x	x	x	x	x	x	x	-	-
I hvilken grad påvirkes du og dine beslutninger av resultatet fra produktet?	x	x	x	x	x	x	x	-	-
I hvilken grad kan du påvirke/påvirker du resultatet fra produktet?	x	x	x	x	(x)	x	x	-	-
I hvilken grad hender det at produktet og din vurdering er forskjellig?	x	x	x	x	-	x	x	-	-
Hvordan bruker du resultatene fra dette produktet/denne modulen?	x	x	x	x	x	x	x	-	-
Hva gjør du om du er usikker på rådet du får fra produktet/modulen?	x	x	x	x	(x)	x	(x)	-	-
I hvilken grad har du, kollegene dine eller andre ved helseforetaket gjort endringer for å tilpasse dere produktet?	x	x	x	x	-	x	x	x	x
Er det annet som du tenker vi bør vite om, relatert til temaene vi har intervjuet deg om?	x	(x)	-	(x)	x	x	x	-	-

Vedlegg 3: Kort om spesialisthelsetjenesten i Norge

Helsetjenestens organisering

Helsetjenesten i Norge er delt i tre nivåer, primærhelsetjenesten (kommunalt ansvar), spesialisthelsetjenesten (statlig ansvar) og tannhelsetjenesten (fylkeskommunalt ansvar). De to casene som belyses i denne studien er begge innenfor norske offentlige sykehus i spesialisthelsetjenesten. Spesialisthelsetjenesten er organisert i fire regionale helseforetak (RHF), Helse Nord RHF, Helse Midt-Norge RHF, Helse Vest RHF og Helse Sør-Øst RHF. RHF-ene eier de 19 offentlige sykehusene og har ansvar for tilbudet i spesialisthelsetjenesten i sin helseregion. I tillegg har hver av de egne sykehusapotek organisert i egne helseforetak og IKT-tjenester organisert i egne IKT-helseforetak (bortsett fra Helse Vest IKT som er et aksjeselskap). Andre oppgaver ivaretas gjennom nasjonale selskaper som regionale helseforetak eier i fellesskap, som Sykehusinnkjøp HF, Pasientreiser HF, Luftambulansetjenesten HF, Helsetjenestens driftsorganisasjon HF og Sykehusbygg HF. *Kilde: Helsepersonellkommisjonen (2023) og Regjeringen (2014).*

Diagnose og behandling

Når helsetjenesten leverer helsetjenester, er det to overordnede prosesser som gjennomføres: Å sette diagnose og å bestemme behandling. Begge består av flere steg hvor data brukes for å gjøre gode beslutninger. Maskinlæring har potensiale til å forbedre hypotesegenerering og hypotesetesting for begge disse (Panch et al., 2018).

Lovregulering av helsetjenester

En faktor som er viktig for hvordan man bruker kunstig intelligens i helsesektoren er at tjenestene, det medisinske utstyret og medikamentene som brukes er regulert i lover. De to enhetene som er caser i denne studien er begge innenfor spesialisthelsetjenesten som reguleres av spesialisthelsetjenesteloven. Prosjekter som forsker på, utvikler eller ønsker å ta i bruk produkter som er basert på kunstig intelligens, må holde seg innenfor eksempelvis helseforskningsloven, helsepersonelloven, regelverket for medisinsk utstyr og personopplysningsloven. Flere av de første

anvendelsesområdene i Norge er innenfor radiologi, og må da være i henhold til strålevernloven og strålevernforskriften (Regelverk og søknader, 2022).

Vedlegg 4: Medisinsk ordliste

Radiologi (generelt)

Computertomografi (CT): I Store medisinske leksikon er CT definert som "en radiologisk undersøkelsesmetode for snittfotografering. I en CT-maskin er det et røntgenrør og diametralt monterte røntgendetektorer, som roterer omkring pasienten under bildeopptak. Røntgenstrålens svekkelse i vevene måles i forskjellige vinkler, blir lagret og behandlet av en computer. Deretter bygges det opp et bilde av de forskjellige vevene i et bestemt snitt eller en bestemt skive"(Brekke et al., 2022).

Brekke et al. skriver videre at moderne CT-maskiner gir svært gode bilder i løpet av sekunder. I tillegg er det ting man kan gjøre for å lette tolkningen: "For å forsterke forskjellene mellom ulike vev kan CT-undersøkelsen gjøres med forskjellige røntgenkontrastmidler. I tillegg til vanlige tverrsnittsbilder kan det lages tredimensjonale fremstillinger av kroppens blodårer (CT-angiografi) og andre organer".

Røntgenlege / radiolog er to betegnelser på samme yrke, og "er en lege med godkjent spesialistutdanning i medisinsk radiologi"(Brekke, 2018). Legeutdanning blir gitt ved fire offentlige universiteter i Norge, og tar 6 år. Spesialisttreningsprogrammet er delt i tre moduler som tas over 6,5 år. Ikke alle spesialitetene krever alle tre modulene. Mens legene spesialiserer seg, blir de kalt leger i spesialisering (LIS) (*LIS*, 2022; Sperre Saunes et al., 2020).

Radiograf: *"En yrkesutøver som utfører ulike bildediagnostiske undersøkelser som røntgen, CT, MR og nukleærmedisinske undersøkelser. Videre assisterer de leger ved invasive undersøkelser og behandlinger (…) Stråleterapi utføres av radiografer med egnet fagkompetanse"(Brekke & Borthne, 2022a)*.

Case 1 Begreper knyttet til segmentering ved doseplanlegging

Ablatio er fjerning av en kroppsdel. I case 1 gjelder dette fjerning av et bryst i forbindelse med kreft. *Kilder: https://kreftforeningen.no/om-kreft/undersokelser/mammografi og https://sml.snl.no/ablatio*

Doseplanlegging av strålebehandling ved kreft: Strålebehandling er medisinsk behandling ved hjelp av ioniserende stråling. Strålebehandling er en viktig del av kreftbehandling, og omfatter forskjellige tekniske metoder for å oppnå høy stråledose i svulstene. Etter kartlegging av kreftsykdommens utbredelse utføres undersøkelse med computertomografi (CT) av aktuelle kroppsområder i den stillingen pasienten skal ha under strålebehandlingen. CT-bildene overføres til et doseplansystem, og legen bestemmer hvilket område i kroppen som skal bestråles og hvilke kombinasjoner av strålefelt som gir best mulig dosefordeling (høy for svulsten (målvolumet), og akseptabelt lav for normale organer som skal skjermes). Kilde: https://sml.snl.no/strålebehandling

Postoperativ av latin post, 'etter', og operatio, 'arbeid'. Det som skjer etter operasjon. Kilde: https://sml.snl.no/postoperativ

Segmentering eller bildesegmentering, betyr å dele opp i segmenter/avgrense deler av et bilde. I case 1 gjøres dette på CT-bilder som en del av planleggingen av strålebehandlingen. Etter at CT-bildene er tatt, tegner man rundt strukturer som enten skal være mål for strålingen ("målvolumet") eller som skal skjermes for strålingen ("risikoorganene").

Serom er en ansamling av sårvæske. Det kan for eksempel opptre i sårhulen etter operasjoner i bryst eller armhule. *Kilde: https://sml.snl.no/serom*

Stråleterapi, strålebehandling er høyenergetisk røntgenstråling mot et område. Dette påvirker kreftcellenes arvemateriale, slik at de enten dør eller slutter å dele seg. På denne måten er det mulig å helbrede og/eller begrense spredningen av kreft. *Kilde: https://kreftforeningen.no/om-kreft/behandling/stralebehandling*

Case 2 Begreper knyttet til koronar angiografi

Hjertet og hjertekamrene: Hjertet er en dobbel pumpe, delt i en venstre og en høyre halvdel. Begge halvdelene har et forkammer og et hovedkammer (også kalt hjertekammer). Hjertet er en dobbel

trykk- og sugepumpe, delt i en venstre og en høyre halvdel. Venstre hjertehalvdel mottar det oksygenrike blodet fra lungene og pumper det ut i kroppen gjennom den store hovedpulsåren, aorta (det store kretsløp). Høyre hjertehalvdel mottar det oksygenfattige blodet fra kroppens organer og pumper det ut i lungene via lungepulsåren (det lille kretsløp), der gassutvekslingen av karbondioksid og oksygen foregår. Hver gang hjertet trekker seg sammen vil trykkstigningen forplante seg i blodet gjennom arteriene. Denne trykkstigningen kalles pulsen. Kilde: https://sml.snl.no/hjertet

Invasiv undersøkelse betyr i denne konteksten at man fører instrumenter inn i blodårene for å undersøke dem innenfra. Dette gjennomføres av en hjertelege, **kardiolog**, som man i intervjuene også omtaler som "invasiv kardiolog".

Koronar angiografi er røntgenundersøkelse av koronararteriene. Et kateter (tynt plastrør) føres fra et innstikk i en pulsåre inn i hovedstammen til koronararteriene, hvorfra det sprøytes inn kontrast og tas bilder.

Faseinndeling av hjerteslagets syklus: Store medisinske leksikon beskriver hjerteslagets syklus slik: *"Hjertet veksler mellom to faser: sammentrekningsfasen (systole) og fyllingsfasen eller hvilefasen (diastole). Disse to fasene kalles til sammen for en hjertesyklus. Vanligvis har vi 60 til 80 hjertesykluser hvert minutt i hvile"* (Arnesen et al., 2022).

For å ta CT-bilder på optimalt tidspunkt innad i hjerteslaget, kobler man på EKG-utstyr og tar en eller flere bildeserier nå nær diastolen som mulig. Ulike leverandører av utstyr deler opp tiden mellom to hjerteslag i såkalte faser. Dersom man deler inn i 20 faser, kalles disse typisk 0%, 5%, 10% og så videre, ifølge IT-radiograf.

Koronararteriene: Hjertemuskulaturen får blodtilførsel gjennom kransarteriene eller koronararteriene. Det er to arterier (arteria coronaria dextra og sinistra) som begge utgår fra den oppadstigende delen av hovedpulsåren (aorta ascendens), like etter avgangen fra venstre hovedkammer (Arnesen et al., 2022).

Vedlegg 5: Sitater og avidentifisering

Tabellen under viser hvordan sitater er håndtert:

Behandling av sitat	**Eksempel (ikke alle er brukt i oppgaven, men alle er i transkripsjonen)**
For personer som bruker mange fyllord har jeg fjernet enkelte typiske fyllord fra sitatene, **der det ikke har hatt noe å si for betydningen**. Jeg har vært nøye med å sjekke at meningsinnholdet ikke ble endret før jeg har tatt bort fyllordene. Eksempel på fyllord som ble brukt en del var *"da"*, *"eh", "(og) så", "sånn"* og *"jo"*	*"mye bedre situasjoner hvor man kommer utenfor boksen, ~~da,~~ utenfor rammene"* "*det er ~~jo~~ hjerte og lunge, thyroidea og esophagus"* *"Selv om det er et veldig bra produkt, ~~så, eh…~~ men det at jeg jobber veldig mye alene"*
Muntlig språk er av og til veldig forskjellig fra skriftlig. Jeg har fjernet deler av setninger hvor personen har stoppet midt i en setning og startet setningen på nytt, både for å gjøre sitatene kortere, men også fordi de blir klarere da og ikke så forvirrende å lese.	*~~"Dette er ikke noe vi har gjort,.. eh…~~ vi har jo ikke stoppet produksjonen, så alt dette er jo arbeid med innsamling er gjort i tillegg til vanlig drift"* *"…en hovedbeslutning er om hun skal ha strålebehandling, for det kan være at ~~hun.. at~~ det i samtale med pasienten dukker opp forhold…"* *"…eller at ~~jeg skriver… altså…~~ det kan være litt motstridende svar"* *"Det er den tilbakemeldingen de trenger, og ~~den …~~ jeg har ikke"*

	"…med tilbakemelding på at de syntes ~~… de trodde at vi gav nitro … eller at~~ vi scannet for kort tid etter at vi gav nitro" *"Det har så mye å si for ~~hvor …~~ utvidelsen av hjertekarene, ~~… hvor …~~ de blir jo større når vi har gitt nitro"*
Noen setninger eller deler av setninger er tatt bort, for å korte det ned uten å miste det vesentlige. Der det er utelatt noe er det markert med "(…)"	*"Deler av setning (…) En annen setning"*
Dersom et sitat består av flere setninger som er sagt rett etter hverandre, er alle setningene innenfor to anførselstegn	*"Dette er setning1. Dette er setning2."*
Dersom det er flere setninger, men enkelte setninger som er sagt i mellom er tatt bort, har de hver sine anførselstegn	*"Dette er setning 1", "dette er setning 3"*
Dersom jeg har lagt til ord for å knytte sammen setningene, så står det jeg har lagt til utenfor anførselstegnene, og det er ikke i kursiv skrifttype	*"Dette er setning 1"*, fortalte vedkommende, og utdyper videre at *"dette er setning 3"*
Jeg har satt ordforklaring i klammer der det ikke fremgår av sitatet hva intervjupersonen viser til (det fremgår et annet sted i transkripsjonen)	*"Du ønsker å gi så høy dose som du kan til det området, og så spare i mest mulig grad"* *=>"Du ønsker å gi så høy dose som du kan til det området, og så spare i mest mulig grad [friskt vev]"*
Der jeg har tatt bort en del inn i sitatet, har jeg erstattet denne delen med parentes med tre	*"La oss si at vi tegner på annethvert bilde, og så gir programmet oss forslag til det bildet vi har*

prikker inni. Dette har jeg bare gjort dersom jeg mener at meningsinnholdet i sitatet der det står i teksten ikke er endret av at jeg har tatt bort en del, men markeringen gjør at man senere kan ettergå hva som var borte	*hoppet over, ~~så vi kan interpolere,… vi tegner halvparten av bildene og~~ så blir det på en måte lagt i hop til en helhetlig struktur til slutt"* *=> "La oss si at vi tegner på annethvert bilde, og så gir programmet oss forslag til det bildet vi har hoppet over (…) så blir det på en måte lagt i hop til en helhetlig struktur til slutt"*

For å ikke få fokus på det konkrete sykehuset og personene der, har jeg gjort enkelte tiltak for avidentifisering teksten. Disse er beskrevet med eksempler her.

Avidentifiseringstiltak i teksten	**Eksempel**
Jeg har skrevet sitatene på bokmål uansett hvilken dialekt eller målform intervjupersonene brukte	*…litt vanskeleg å svara på => litt vanskelig å svare på*
Jeg har anonymisert sykehuset og avdelingen	Eksempler: Sykehus 1 Enhet 2
Uansett kjønn, så har jeg omtalt personen som "han" eller "vedkommende"	Hun => han, vedkommende