NTNU
Det skapende universitet

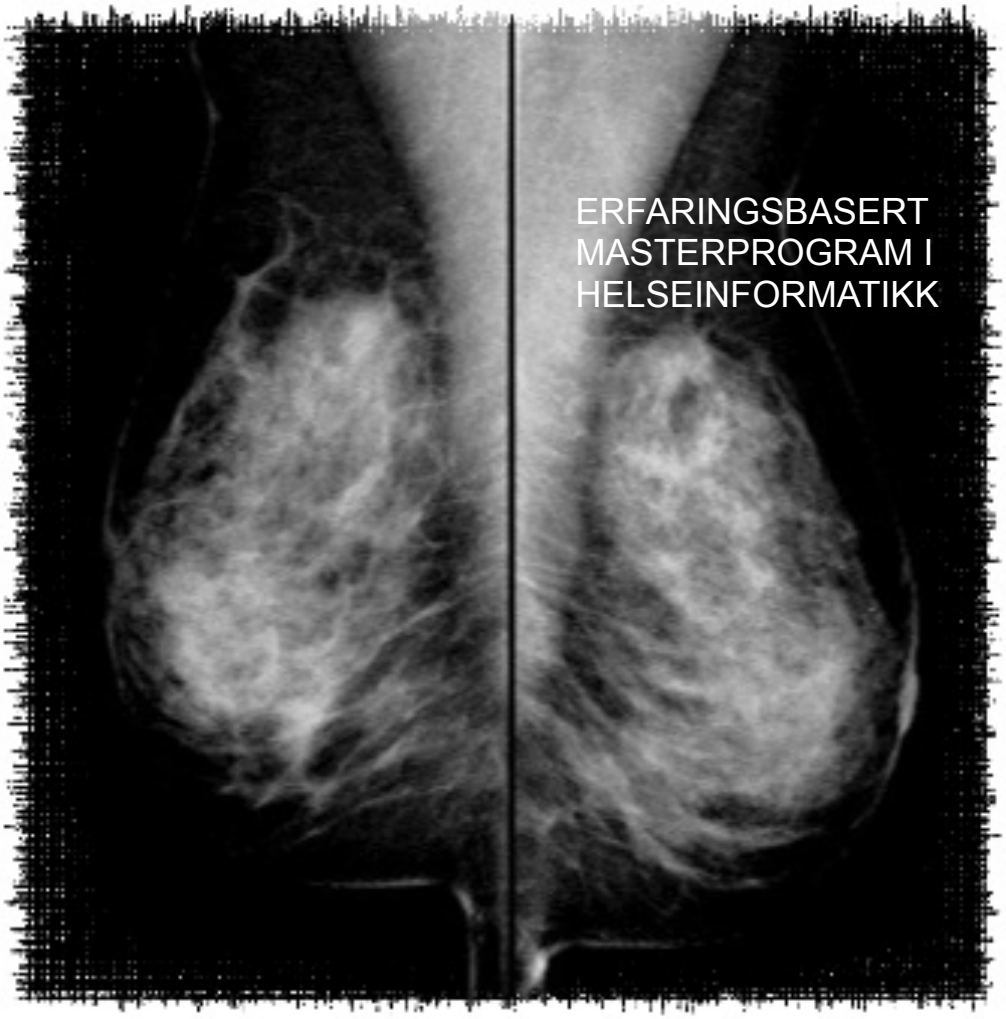
Etablering av testsett for radiologer som tyder mammografibilder

Arvid Austgulen
Anne Kathrin Olsen Ertzaas

Helseinformatikk
Innlevert: Desember 2012
Hovedveileder: Pieter Jelle Toussaint, IDI
Medveileder: Gunnar Klein, IDI
Solveig Hofvind, Kreftregisteret

Norges teknisk-naturvitenskapelige universitet
Institutt for datateknikk og informasjonsvitenskap

MDV 6191 MASTEROPPGAVE
Fakultet for datateknikk og informasjonsvitenskap



Etablering av testsett for radiologer som tyder mammografibilder

Anne Kathrin Ertzaas
Arvid Austgulen
2012

MDV 6191 Masteroppgave

Erfaringsbasert masterprogram i helseinformatikk, NTNU

Etablering av testsett for radiologer som tyder mammografibilder

Anne Kathrin Ertzaas
Arvid Austgulen

---

12. desember 2012

Forord

Dette dokumentet samt utviklet e-læringssystem inklusive screening-mammogrammer for tyding av testsett utgjør en masteroppgave i emnet MDV6191. Oppgaven utføres i 7. og 8 semester i erfaringsbasert masterprogram i Helseinformatikk ved NTNU Videre i Trondheim. Studietilbudet er et samarbeid mellom Det medisinske fakultet (DMF), Institutt for datateknikk og informasjonsvitenskap (IDI) og NTNU Videre.

Begge studenter Arvid Austgulen, Curato Røntgen, Bergen og Anne Kathrin Ertzaas, Kreftregisteret, Oslo har helsefaglig bakgrunn.

Tre professorer har vært veiledere for vårt arbeid: Pieter Toussaint ved IDI NTNU, Gunnar Klein ved NSEP NTNU og Solveig Hofvind ved Kreftregisteret og HiOA.

Det er mange som skal takkes ved avslutningen av et masterprosjekt. Først og fremst vil vi takke veilederne som hver og en ga oss gode råd, innspill, korreksjoner og spennende diskusjoner i ulike deler av arbeidet.

Takk til:

- radiologene som stilte opp og delte av sin kunnskap, kom med innspill, krav, ønsker og ga ris og ros
    - en spesiell takk til Per Skaane, Ullevål Universitetssykehus som har re-gransket mammogrammene og testet programvaren
    - de fire radiologene som har testet et ferdig testsett
- radiograf- og radiologledere ved de fire BDS vi har hentet bilder fra: for velvillighet, tilrettelegging og hjelp ved uthenting av mammogrammer fra sykehus PACS`ene. En spesiell takk til Berit Hanestad, Haukeland Universitetssykehus som hentet ut bilder for oss.
- kollegaer som har bidratt med gode innspill, gjennomlesing og konstruktiv kritikk

- våre arbeidsgivere som har lagt til rette for at vi har kunnet gjennomføre studiene
- våre familier for tålmodighet, oppmuntring og støtte i forbindelse med studier og masteroppgaven

Arbeidet med masteroppgaven har vært en utviklende, arbeidskrevende og spennende læringsprosess. Vi har hatt et velfungerende samarbeid og en god fordeling av arbeidsoppgaver. Programmeringen er utført av Arvid. Den skriftelige delen av masteroppgaven er i hovedsak ført i pennen av Anka.

Desember 2012

Arvid Austgulen, Bergen

Anne Kathrin (Anka) Ertzaas, Lørenskog

Innholdsfortegnelse

Liste over tabeller og figurer

Sammendrag

Effekten av mammografiscreeningprogrammer er avhengig av at radiologene har høy kompetanse. Et sentralt spørsmål er hvordan vedlikeholde og videreutvikle radiologenes tydekompetanse. Å benytte testsett for egentesting kan være et bidrag til å øke sensitiviteten på tyding av screeningmammogrammer, som et supplement til anbefalt kvalitetssikringsarbeid.

Vi har etablert systemet OPTIMA som består av programvare for tyding av screeningmammogrammer, inklusive fire testsett bestående av 100 screeningundersøkelser. Systemet har et registreringssystem og en rapportmodul.

Hvert testsett består av 100 screeningundersøkelser med mammogrammer fra én utstyrsleverandør; General Electric, Hologic, Philips Sectra eller Siemens. Testsettene består av sanne negative (85-75%) og sanne positive screeningundersøkelser (15-25%), tilfeldig valgt etter et randomisert utrekk fra den nasjonale mammografidatabasen. Mammogrammene er hentet fra fire ulike sykehus PACS og deretter anonymisert.

Systemet samler informasjon om tydernes ferdigheter ved å benytte modifisert BI-RADS klassifisering benyttes for registrering av selekterte funn. OPTIMA gir umiddelbar tilbakemelding på testers ferdigheter sammenlignet med ”fasit”. Systemet gir mulighet for å gå tilbake og vurdere mammogrammene på nytt der eget tyderesultat eller BI-RADS klassifisering er i uoverensstemmelse med ”fasit”. Systemet gir mulighet for å benytte zoom og window/level i tydeprosessen.

Radiologer bidro i planlegging, testing og ferdigstilling av systemet. Brukermedvirkning har vært hensiktsmessig for å utvikle og forbedre våre kravspesifikasjoner for sammensetning av testsettene, programvaren og innhold i registreringsløsningen. Detaljerte brukerkrav ble etablert for sikre at systemet var i henhold til radiologenes ønsker og behov. Gode brukerkrav og innspill har resultert i

et brukervennlig system, noe radiologene bekrefter ved å gi programvaren høy score i brukervennlighetstesting.

Fire radiologer har tydet ett testsett hver. Resultatene viser liten variasjon i seleksjon av de sanne positive funnene, men det var interobservatørvariasjon i hvordan radiologene klassifiserte BI-RADS for sanne positive funn. Vi mener testsettene kan stimulere til læring og kompetanseutvikling for radiologers tyding av screening-mammogrammer, spesielt for å redusere variasjonen i BI-RADS klassifisering av funn.

Summary

The effectiveness of mammographic screening depends on the radiologist's ability to interpret the screening mammograms. A key issue is to maintain and improve the interpretation skills. One way to do this is to let the radiologists participate in a regular educational self-assessment scheme. Test sets of screening mammograms might be one way to increase the sensitivity of the reader performance as a supplement to the recommended quality assurance.

We have created a software system called OPTIMA for interpreting screening mammograms. OPTIMA include four test sets of screening mammograms, a registration system and a report module.

Each set of 100 screening mammograms contains mammograms from one supplier; General Electric, Hologic, Philips Sectra or Siemens. Screening examinations including true negative (85-75%) and true positive cases (15-25%) are randomly selected from the national screening database. All mammograms are obtained from a hospital PACS and thereafter anonymoused.

The system collects information about the radiologist interpretation of the screening mammograms by using modified BI-RADS classification of the mammographic findings of positive cases. OPTIMA include a report module to give immediate

feedback of the interpretation and classification. It is possible to return to mammograms with incorrect interpretation and/or classification. There are options related to zoom, window and level during interpretation.

Radiologists contributed in the planning, testing and finalizing the system. We established detailed requirements to ensure that the software-system is in accordance with radiologists needs. The radiologists validate software usability by using a "System Usability Scale“.

Four radiologists interpreted the test sets. The results show little variation in the selection of true positive findings. However, there was a variation in how the radiologists used the modified BI-RADS classification for breast cancer cases. The test set will work as an educational self-assessment and training scheme for radiologists, particularly as a tool to reduce the variability in BI-RADS classifications for mammographic findings.

1 Innledning

Mammografiscreening er basert på en røntgenologisk prosedyre, og mammografibildene vurderes (tydes) av en radiolog. Både den norske og europeiske kvalitetsmanualen for organisert mammografiscreening anbefaler at radiologer som arbeider i mammografiscreening skal tyde minst 5000 screeningundersøkelser årlig, i tillegg til å arbeide med klinisk mammografi, ultralyd, MR og være øvet i trippeldiagnostikk (1,2). Dobbel uavhengig tyding av screeningmammogrammene er anbefalt (1,2).

Mammografiprogrammet er det offentlige mammografiscreeningprogrammet i Norge som inviterer kvinner i alderen 50-69 år til tobilde screeningundersøkelse hvert annet år. Kreftregisteret administrerer programmet og har databaser med oversikt over alle opplysninger knyttet til screeningprosessen. Mammografiprogrammet benytter utelukkende radiologer som screenere (tydere) av mammogrammene, hvilket også er tilfelle i de fleste land og kontinenter (1,2). For å oppnå anbefalte kvalitetsmål (tidligindikatorer) er det viktig at radiologene har høy faglig kompetanse. Selvevaluering er et bærende element i kvalitetssikring i Mammografiprogrammet (1). Det er etablert retningslinjer som har til hensikt å fremme kvalitet, forbedre arbeidsrutiner og redusere antall feil i arbeidet (1). Regelmessig egentesting ved å benytte testsett med screeningmammogrammer kan være et virkemiddel i å fremme læring og et viktig verktøy for kvalitetssikring og kvalitetsutvikling av tydeprosessen (3,4,5,6,7). Et program og testsett kan brukes som en av mange etablerte opplæringsmetoder og til å teste individuelle ferdigheter knyttet til tydeprosessen av screeningbilder. Slike testsett kan være sammensatt av screeningmammogrammer med ulike funn; normale, suspekte funn og klare malignitetssuspekte forandringer.

Vi ønsker å etablere et system som et bidrag til å opprettholde og heve kompetanse knyttet til interpretasjon og persepsjon av mammografi. I tillegg vil det kunne bidra til å opprettholde høy kvalitet på det radiologfaglige arbeidet i Mammografiprogrammet.

Programvaren vil ha en rapportmodul som gjør at radiologene får umiddelbar tilbakemelding på testen.

1.1 Oppgavens omfang og avgrensning

Denne rapporten er sammen med utviklet programvare vår masteroppgave i studiet helseinformatikk i faget MDV6191 ved NTNU. Dokumentasjonen av kravutviklingen er beskrevet i metode- og resultatkapittelet.

Rapporten inneholder teori om screening, brystkreftepidemiologi og Mammografiprogrammet, radiologisk virksomhet i mammografiscreening, metoder for evaluering av diagnostiske tester, hvordan kompetanse etableres og hvordan testsett av mammogrammer kan benyttes som e-læringsverktøy. Rapporten inneholder beskrivelse av metode for identifikasjon av screeningmammogrammer og kravutvikling for programvare. Og rapporten inneholder resultater som er systemspesifiksjonen og validering, samt diskusjon for å besvare oppgavens problemstillinger.

1.1.1 Målsetting for masteroppgaven

Målsettingen med masteroppgaven er å etablere testsett med screeningmammogrammer samt utvikle programvare for tyding bildene. Programmet bør umiddelbart gi resultater etter gjennomført tydetest. Målsettingen er å bidra til å opprettholde og heve kompetanse knyttet til interpretasjon og variabiliteter knyttet til persepsjon av mammografi. Vi ønsker programvaren med testsettene kan stimulere interesse for tyding av mammogrammer og være et verktøy som radiologene kan benytte for å teste individuelle ferdigheter i tydeprosessen og som en opplæringsmetode innen mammadiagnostikk.

Masteroppgaven består av flere praktiske deloppgaver:

1. Identifikasjon av et randomisert utvalg av screeningmammogrammer fra Mammografidatabasen ut fra gitte kriterier
2. Innhente screeningmammogrammer fra fire sykehus PACS og sammensetning av testsettene

3. Utvikle programvare for tyding av mammogrammene og klassifisering av positive funn i testsettene
4. Brukbarhetstesting av programvare
5. Utvikle et tilbakemeldingssystem for gjennomført tyding av testsettene
6. Utprøving av testsettene

Oppgave er primært et kravutviklingsprosjekt som krever en bred teoretisk forankring. Siden vi skal etablere programvare for tyding av testsett vil utvikling av programvaren være en del av masteroppgavens metode. Metodedelen vil derfor beskrive faser av utviklingsarbeidet med brukerstyrt utvikling og brukbarhetstesting av en programvare-prototype. Dokumentasjon av kravspesifikasjonene og programvaren utgjør resultatdelen av oppgaven.

1.2 Problemstillinger

Målsettingen med testsettene er å etablere et utvalg screeningmammogrammer som kan reflektere en normal screening- og tydesituasjon. Testsettene vil bestå av mammogrammer fra en screeningundersøkelse. Testsettene inneholder kun sanne negative (SN) og sanne positive (SP) screeningfunn.

Oppgaven vil ha en teoretisk og en praktisk tilnærming og problemstillingene er som følger:

- Hvilke læringsteorier bør være grunnlag for å utvikle programvare for tyding av testsett med screeningmammogrammer?
- Hvilke kravspesifikasjoner kreves for å utvikle programvare for tyding av testsett og sammensetning av testsettene?
- Hvilke praktiske forhold bør det tas hensyn til ved utvikling av et system for tyding av screeningmammogrammer?
- Vil brukermedvirkning øke programmets anvendbarhet og brukbarhet?

1.2.2. Begrensninger ved oppgaven

Begge studenter har helsefaglig bakgrunn, noe som kan være en begrensende faktor i utvikling av programvare. Prosjektet har ikke hatt økonomisk støtte. Dette begrenser valg av testpersoner og uthenting av bilder, rent geografisk.

Masteroppgaven fokuserer på å teoretisere, kravspesifisere og programmere programvare for testsett. Programvare for testsettene er ferdigstilt, men ikke distribuert til alle landets BDS innenfor vår prosjektperiode.

1.2.3 Definisjoner og terminologi

Det benyttes ord og utrykk både fra medisin og helseinformatikk. Disse er beskrevet i tabellen under.

Tabell 1: Ord og utrykk med forklaring

Ord/uttrykk	**Forklaring**
Arbitration-cancere	Screeningoppdaget brystkreft, selektert til konsensus (diskusjon) av kun en av de to tyderne
BI-RADS	Breast Imaging Reporting and Database System: kodeverk utarbeidet for å standardisere brystdiagnostikk og fasilitere overvåking av utfallet
Brukskvalitet (usability)	Den opplevde kvaliteten av et produkt i bruk, i denne oppgaven av radiologer
Brystkreft	Begrepet brukes om både Ductalt carsinom In Situ = DCIS (malign svulst type 2) og infiltrerende brystkreft/cancer (malign svulst type 3)
DCIS	Ductalt carsinom In Situ; kreft lokalisert i melkegangene (pre-invasiv/ikke - infiltrerende brystkreft)
Deteksjonsrate	Rate av kvinner diagnostisert med brystkreft per 1000 screenede. Deteksjonsraten kan være for DCIS og for infiltrerende brystkreft og for DCIS og infiltrerende samlet
Dismissed	Funn på mammogrammer som etter diskusjon på konsensusmøtet deselekteres og kvinnen blir ikke innkalt til etterundersøkelse
HPR	Helsepersonellregisteret er helsemyndighetenes register over alt helsepersonell med autorisasjon eller lisens etter helsepersonelloven
Hengeprotokoll	Ønsket visning av screeningbildene. Eksempel: Høyre skrå bildevises øverst til venstre, venstre skråbilde vises i øverst til høyre; høyre frontbilde i nedre venstre og venstre frontbilde nedre høyre
Insidens	Forekomst: Antall nye tilfeller av sykdom i en spesifikk

	populasjon i løpet av et gitt tidsrom
Insidensrate	Antall syke/ personår (per 100 000)
Intervallkreft	Brystkreft (DCIS eller infiltrerende) detektert etter en negativ screeningundersøkelse, med eller uten tilleggsbilder og nåleprøve, før neste planlagte screeningundersøkelse
Kohort	Gruppe som har en felles faktor. En statistisk term som blir brukt ved inndeling av et undersøkelsesmateriale, eks. kvinner født i 1955
Mammografiprogrammet	Det offentlige screeningprogrammet i Norge, som inviterer kvinner i målgruppen 50-69 år til mammografi hvert 2. år.
Mortalitet	Dødelighet: Antall som dør av en gitt sykdom i en spesifikk populasjon i løpet av et gitt tidsrom
OPTIMA	Et e-lærinsgssystem som inneholder programvare og mammogrammer for tyding
PACS	Picture Archiving and Communication System - elektronisk bildarkiv som benyttes for å lagre røntgenbilder ved helseforetakene
PPV	Positiv prediktiv verdi: Andel brystkrefttilfeller blant kvinner som er innkalt til tilleggsundersøkelse på grunn av mammografifunn
Prevalens	Andel med sykdom i en spesifikk populasjon på et gitt tidspunkt (for brystkreft; andel kvinner som lever med sykdommen)
Prevalent screenet	Første screeningundersøkelse i Mammografiprogrammet
RIS	Radiologisk informasjonssystem
ROC	Receiver Operating Characteristic. ROC kurve er en grafisk fremstilling av sensitivitet vs spesifisitet
Sensitivitet	Testens evne til å påvise sykdom, det vil si å diagnostisere de reelt positive
Spesifisitet	Sannsynlighet for å ha negativ test hvis man er frisk dvs. testens evne til å oppdage de reelt negative
STARD	Standards for Reporting of Diagnostic Accuracy - Standarder for rapportering av diagnostisk nøyaktighet
Subsekvent screenet	En påfølgende screeningundersøkelse. Tidligere undersøkelse i programmet er for to eller flere år siden.
Tidligindikatorer	Overordnede kvalitetsparametre for screeningprogrammet (surrogat- eller delmål)

Tyder	Radiolog som vurderer mammogrammene
UML (The Unified Modelling “Language”)	Standardisert grafisk notasjon for å støtte objektorientert analyse og utforming
Validering	Vurdering av om systemet tilfredsstiller ønsker og at utførelsen tilfredsstiller krav
Window and Level	Et interaktivt verktøy for kontrastforbedring benyttet innen bildebehandling

1.2.4 Gjennomføring av arbeidet

Vi har valgt en praktisk tilnærming til arbeidsmøter og veiledning da studenter og veileder er bosatt i ulike deler av landet. Studentene er bosatt i Bergen og Oslo. To veiledere befinner seg i Trondheim og en i Oslo. Vi har brukt DropBox for å dele manuskript og programmert materiale, med egne mapper for studenter og veiledere. Studenter og to av veilederne har etter avtale hatt Skype- og telefonmøter om fremdrift av og veiledning i arbeidet. Studenter og en veileder møttes våren 2012 i Oslo hver annen uke i forbindelse med jobbreise. Notat i forkant og referat i etterkant av veiledning har dokumentert fremdrift og gjøremål i arbeidet.

2 Teori

Mammografiscreening er basert på en røntgenologisk prosedyre. I mammografiscreening tas det primært to bilder av hvert bryst. Bildene screenes uavhengig av to radiologer. Effekten av screeningprogrammer er dermed avhengig av at radiologene har en høy faglig kompetanse innen screeningmammografi. Det er etablerte mål for radiologisk virksomhet i mammografiscreening i norsk og europeisk sammenheng (1,2). Radiologenes kompetanse opparbeides gjennom spesialistutdanning og erfaring.

2.1 Screening

Masseundersøkelser mot kreft (screening) har til hensikt å oppdage krefttilfeller eller forstadier til kreft før sykdommen har spredt seg (8). Målsetning med tidlig diagnostisering er å kunne iverksette kurative tiltak så tidlig som mulig, eller gjennom tidlig behandling endre sykdomsforløpet. Da har behandlingen bedre forutsetninger for å lykkes. Screening og tidlig diagnostisering kan rettes mot den generelle befolkningen eller personer med antatt høyere risiko for å utvikle sykdom. Screening som metode benyttes oftest som ledd i forebyggende helseprogram der målet (endepunktet) er å redusere sykdom og død. Forutsetninger i et screeningprogram er at sykdommen må gi alvorlige helseproblemer, har en relativt lang symptomfri fase, ikke bør være en sjelden sykdom og screeningtesten må være pålitelig. Grunnlaget for å tenke screening er at det eksisterer en screeningtest som er i stand til å indikere aktuell sykdom, at testen er enkel, akseptert av befolkningen og er kostnadseffektiv. WHO beskriver ti punkter som grunnlag for å etablere et screeningprogram for en sykdom (8)

Mammografiscreening er beheftet med fordeler og ulemper. Ulemper ved screening kan være økt insidens og prevalens, falske positive funn, uro ved etterinnkalling (9,10,11,12), overdiagnose (13,14), overbehandling og intervallkreft (15). Fordeler nevnes som redusert dødelighet, stadieforskyvning ved diagnose, mindre omfattende

behandling og økt tverrfaglig kunnskap blant annet ved etablering av brystdiagnostiske sentra (15,16,17). Det er bred internasjonal enighet om at mammografiscreening reduserer dødeligheten av brystkreft (18,19,20,21,22,24). Reduksjonen er 25 % og 31 % blant de som inviteres til screening ("intention to treat"), mens den er mellom 38 % og 48 % blant de som faktisk deltar i programmet. Effekten er mest overbevisende blant kvinner 50 – 69 år og har blitt vist i randomiserte kontrollerte forsøk (18,19,20,21,22,23,24). Fordelene er større enn ulempene ved organisert mammografiscreening (25). Det er ikke innenfor denne oppgavens rammer å diskutere fordeler og ulemper ytterligere.

2.1.1 Mammografiprogrammet

På bakgrunn av prøveprosjekt med masseundersøkelse for brystkreft med mammografi vedtok Stortinget i 1998 å innføre mammografiscreening som et landsdekkende program, Mammografiprogrammet, som en del av det offentlige helsetjenestetilbudet (26). I 2001 ble Kreftregisteret utpekt av Sosial- og helsedepartementet til å være et nasjonalt screeningsenter i Norge. Per i dag har Kreftregisteret ansvar for drift av nasjonale screeningprogrammet mot brystkreft, livmorhalskreft og et forprosjekt med tarmkreftscreening (27).

Mammografiprogrammet inviterer alle kvinner i aldersgruppen 50-69 år til screeningundersøkelse hvert annet år. Målgruppen i 2012 er 565 000 kvinner. Målsettingen med programmet er å redusere dødeligheten blant de inviterte. Siden et slikt endepunkt kan måles først mange år etter oppstart blir programmet til enhver tid målt for en rekke prosessindikatorer eller delmål.

All informasjon knyttet til kvinnens screeningundersøkelse fra invitasjon, oppmøte, tyderergistreringer og eventuelle etterundersøkelser er samlet i en felles database på Kreftregisteret. Det er publisert omlag 120 vitenskaplige artikler fra data fra den nasjonale screeningdatabasen med ulike aspekter fra programmet (9,10,11,12,13,14,15,16,17,18,28,29,30). Ytterligere ekstern evalueringen av ulike

aspekter av Mammografiprogrammet som skal gjennomføres av 7 forskningsgrupper i regi av Norsk Forskningsråd. Evalueringen startet våren 2012.

2.1.2 Brystkreft i Norge

For å kunne teste for en sykdom må en ha kunnskap om sykdommen ved å studere dens forekomst relatert til ulike dimensjoner. Disse kan være tid (forekomsten og eventuell endring over en viss periode), sted (geografiske forskjeller) og personer (er forekomsten knyttet til spesielle egenskaper eller eksponeringer som personer har eller utsettes for). Noen mål på sykdom og død er prevalens, insidens og mortalitet (31). Brystkreft er den hyppigste kreftformen blant kvinner i Norge. I 2010 fikk 2839 kvinner diagnostisert sykdommen (32). I det samme året døde 673 kvinner av sykdommen. Forekomsten av brystkreft har økt jevnt fra 50-tallet, med en utpreget økning i perioden hvor den offentlige mammografiscreeningen i Norge startet i 1996-97. Overlevelse av brystkreft er nært relatert til histopatologisk tumorkarakteristikk og av stadium ved diagnose. Kreftregisteret registreres opplysninger om tumorstørrelse, -grad, lymfeknute- og hormonreseptorstatus og ulike typer behandling. Dette gjøres for å analysere brystkreftutviklingen over tid og å relatere informasjonen til overlevelse og dødelighet samt for å vurdere ulike behandlingsopplegg for kvinnene.

2.2 Evaluering av diagnostiske tester

Systematiske oversikter av diagnostiske tester oppsummerer egenskaper ved testens ytelse men beskriver ikke utfallet for pasienten (33). Sammenhengen mellom testegenskap og pasientutfall er kompleks. Den ultimate hensikten med diagnostiske tester måles ved om resultatet av testen påvirker utfallet for pasienten og som dermed påvirker endepunkt som morbiditet, dødelighet eller livskvalitetsparametre (34). Diagnostiske tester kan påvirke utfallet direkte, men oftest er påvirkningen indirekte. Rent prinsipielt påvirker testresultatet behandlingsmessige valg som eventuelt vil innvirke på resultatet for pasienten. Testens ytelse eller diagnostiske nøyaktighet kan uttrykkes som sensitivitet, spesifisitet, positiv negativ ratio, positiv prediktiv verdi

(PPV), Receiver Operating Characteristic (ROC) og er ofte surrogatmål på vei mot et endepunkt. De fleste diagnostiske tester er beheftet med styrker og svakheter.

Det er viktige prinsipielle forskjeller mellom screening- og diagnostiske tester. En diagnostisk test har som formål å klarlegge en situasjon eller tilstand hos den som blir testet. Screeningtester er undersøkelser for å "sile ut" hvem som må gjennomgå ytterligere tester for å stille en sikker diagnose. Screening kan defineres som "undersøkelse av asymptomatiske individer for å klassifisere dem som sannsynlig friske eller sannsynlig syke i forhold til den sykdommen de undersøkes for" (8). Sensitivitet, spesifisitet og prediktiv verdi er velkjente mål i et screeningprogram og har anbefalte verdier for mammografiscreening. Enhetlig og komplett rapportering er essensielt for å kunne vurdere sensitivitet og spesifisitet for ulike fylker/områder og radiologenes ferdigheter i et mammografiscreeningprogram. Resultatene kan betraktes som et mål på radiologenes ferdigheter.

2.2.1 Sensitivitet, spesifisitet, PPV og NPV

Sensitivitet og spesifisitet gir utrykk for testens nøyaktighet (31). Diagnostisk sensitivitet defineres som sannsynligheten for at testen er positiv hvis man er syk dvs. testens evne til å oppdage de reelt positive. Sensitivitet regnes ut ved å dele sann positiv med sann positiv pluss falsk negativ (SP/SP+FN). Lav sensitivitet betyr at testen gir falsk opplysning om at kvinnen er frisk. Diagnostisk spesifisitet defineres som sannsynlighet for å ha negativ test hvis man er frisk dvs. testens oppdage de reelt negative. Spesifisiteten beregnes ved å dele sanne negative på sanne negative og falske positive (SN/SN+FP). Lav spesifisitet betyr at mange friske får beskjed om at de muligens er syke. Spesifisitet og sensitivitet forteller noe om hvor nøyaktig testen er, det vil si dens evne til å gi rett svar på om den som testes faktisk er eller ikke er bærer av tilstanden det testes på.

En prediktiv verdi gir uttrykk for andelen syke blant de friske. Prediktiv verdi av en test avhenger av eller varierer med forekomsten av aktuell sykdom (prevalens) i den

befolkning pasienten kommer fra. Prediktiv verdi angir sannsynligheten for at en person med utslag på testen (positiv test) virkelig er syk eller virkelig har brystkreft (positiv prediktiv verdi (SP/FN+SP)), eller om sannsynligheten for at en person uten utslag på testen virkelig er frisk (negativ prediktiv verdi (SN/SN + FN)).

Tabell 2 viser hvordan sensitivitet og spesifisitet kan eksemplifiseres for mammografiscreening. Gitt en gruppe på 10000 personer, 100 syke og 9900 friske. Mammografi identifiserer 96 syke som positive (SP), men også 594 friske som positiv (FP). Testen angir 9390 friske som negative (SN), men også 4 syke som negative (FN).

Tabell 2: Screening test for brystkreft

		Virkelig forekomst av brystkreft		Antall totalt
		Brysttkreft JA	Brystkreft NEI	
Prøveresultat (screeningtest)	Positiv +	96 (SP)	594 (FP)	690 (SP+FP)
	Negativ -	4 (FN)	9390 (SN)	9310 (FN+SN)
		100 (SP+FN)	9900 (FP+SN)	10000 (alle)

Sensitivitet = SP/(SP+FN) = 96/100 = 0,96. Mammografiscreeningtestens evne til å påvise sykdom er 96 %.
Spesifisitet = SN/(SN+FP) = 9390/9900 = 0,95. Mammografiscreeningtestens evne til å utelukke friske er 95 %.
PPV = a/(a+b) = 96/690 = 0,13. Dette betyr at 13 % av de syke fanges opp og 87 % er falsk positive. Prevalens = (a+c)/(a+b+c+d) = 100/10000 = 1/100 eller 1 %.

For diagnostiske tester ønskes høyest mulig sensitivitet og spesifisitet; vi ønsker at verdiene er så nær 1 som mulig. Prediktiv verdi av en screeningmetode er avhengig av metodens sensitivitet og spesifisitet samt den aktuelle tilstandens prevalens i den undersøkte befolkningen. PPV for mammografiscreening vil naturlig nok være lav i

forhold til PPV av diagnostisk mammografi. Det vil i en populasjon være delvis overlappende verdier mellom ”friske” og ”syke”, og overlappingen blir større jo dårligere presisjon og reproduserbarhet metoden har. For å optimalisere fordeler og redusere ulemper kan screeningprogram etablere grenseverdier for ønsket og akseptabel etterinnkallingsrate, PPV og deteksjonsrate (1).

2.3 Radiologisk virksomhet i mammografiscreening

For å nå målet om reduserte dødelighet og sykelighet av brystkreft ved å detektere sykdommen i et tidlig stadium kreves nitidig arbeid fra en rekke profesjoner. Siden mammografiscreening er en radiologisk prosedyre avhenger mye av effekten av radiologenes kompetanse, deres evne til å oppdage brystkreft samt korrekt tolking av funn.

Det norske screeningprogrammet gjennomføres etter retningslinjer beskrevet i kvalitetsmanualen (1), som er tuftet på den europeiske kvalitetsmanualen (2). Nasjonale retningslinjer er ment som et hjelpemiddel for å oppnå forsvarlighet og god kvalitet i tjenesten (1). Et screeningprogram må til enhver tid driftes etter gjeldende retningslinjer ved å oppnå kvalitet og for å identifisere forbedringspotensiale. Både den norske og europeisk kvalitetsmanualen for organisert mammografiscreening gir akseptable og ønskede mål for en rekke kvalitetskontrollparametre for virksomheten: Indikatorer er oppmøte, etterinnkallingsrate, PPV, deteksjonsrate, stadium ved diagnose, tumorkarakteristika og intervallkreft (1,2). Flere av indikatorene forteller noe om testens effektivitet, som kan betraktes som et utrykk for radiologen ferdigheter. Et nasjonalt program, brystsenter eller enkeltperson bør tidligst mulig bli oppmerksom på mangelfull kvalitet slik at nødvendige forbedringstiltak kan iverksettes. Løpende evaluering og kvalitetssikring er derfor et krav. Det bærende kravet innen kvalitetssikring er selvevaluering. Kvalitetsmanualen i Mammografiprogrammet har som mål å fremme kvalitet, forbedre arbeidsrutiner og redusere antall feil (1).

Ved oppstart av Mammografiprogrammet ble det etablert brystdiagnostiske sentre (BDS). BDS er regionens senter for brystdiagnostikk der screening- og diagnostiske mammogrammer tydes og lagres. Her foregår etterundersøkelser, samt eventuell videre diagnostikk og behandling. Hvert BDS er tilknyttet en eller flere bildetakingsenheter, der screeningundersøkelsene finner sted. Norge har 16 BDS. Disse er tverrfaglige enhetene er effektiv bruk av spesialisthelsetjenesten (35) gode arenaer for helsepersonells læring (36,37). Et slikt samarbeid er av betydning for å etablere, opprettholde og heve kvalitet og spisskompetanse, både innen brystdiagnostikk og behandling (1,2).

I Norge benyttes utelukkende radiologer som tydere av screeningundersøkelser, hvilket også er tilfelle i de fleste land og kontinenter. Retningslinjer for det radiologiske arbeidet påpeker nødvendigheten av faglighet og tverrfaglighet i arbeidet som utføres ved BDS (1). Det er viktig at radiologene er faglig gode for at anbefalte tidligindikatorer oppfylles og for å nå Mammografiprogrammets mål om redusert dødelighet av brystkreft blant de inviterte. Erfaring har vist at det kan være vanskelig å oppfylle alle anbefalingene (15,16).

Et krav til radiologer som arbeider i Mammografiprogrammet er at de også skal jobbe med klinisk mammografi og ultralyd samt være øvet i trippeldiagnostikk. Å arbeide med mammografiscreening må således sees sammen med en klinisk kontekst ved å forholde seg til videre utredning og funn og krever tverrfaglighet i arbeidet (1). Videre bør radiologen kunne vurdere bruk av MR-mamma og ha tilgang til kliniske opplysninger og patologidata for å bygge opp og øke egen kompetanse. Et annet krav er at radiologer bør ha gjennomført Legeforeningens grunnkurs i mammadiagnostikk eller ha tilsvarende kompetanse og ha deltatt i et internasjonalt kurs i mammadiagnostikk. Det er krav til at radiologer som skal arbeide i Mammografiprogrammet må få individuelt tilpasset opplæring. Det må være etablerte arenaer for tverrfaglighet, kunnskapsformidling og læring (37) for å etablere medisinsk faglig kunnskap og det krever et stort tydevolum for å oppnå og opprettholde

kompetanse. Norske og europeiske retningslinjer anbefaler også at det skal tydes et visst volum screeningundersøkelser årlig (5000 undersøkelser). Læringskurven er brattest de første årene en tyder screeningbilder (38,39). Det har vist seg å være nødvendig av å fokusere på trening/øvelse før oppstart av bildetyding for å øke kompetansen i tydeprosessen for å tilfredsstille kvalitetskrav, spesielt ved overgang fra analoge til digitale tydeprosesser (40,41). Radiologenes erfaring og tydevolum påvirker risikoen for en falsk positiv screeningundersøkelse (42).

Dobbel uavhengig tyding av screeningmammogrammer anbefales og er vanlig praksis i Mammografiprogrammet (1). Dobbel uavhengig tyding med konsensus øker deteksjonsraten sammenlignet med om en radiolog hadde tydet screeningbildene (40). Tydingen av screeningbilder krever mindre rapportering og beskrivelser enn diagnostiske tester. I Mammografiprogrammet benyttes en femdelt tydeskala for å angi resultatet av screeningtesten, standardisert for alle screeningundersøkelser. Tydescore 1 indikerer negativt/normalt funn, 2 sannsynlig benignt funn, 3 usikkert funn, 4 sannsynlig malignt funn og 5 malignt funn. Ulik tydescore (2 eller høyere) fører til at undersøkelsen vurderes av to eller flere radiologer i konsensus, før endelig resultat settes. Om en av radiologene klassifisert undersøkelsen til 3, 4 eller 5 skal kvinnen etterinnkalles til undersøkelse. Dette kan fravikes i en opplæringssituasjon, og er et viktig læringsmoment i opplæring av nye radiologer.

En norsk studie viser at 24 % av screeningoppdagede krefttilfeller ble selektert av kun av en av de to radiologene (28). Det er anbefalt å inkludere foregående undersøkelse for sammenligning og at mammografibilder fra tidligere undersøkelser er tilgjengelig (1,44).

Radiologen som tyder screeningundersøkelser kan identifiseres og resultatene følges på individnivå, både ved det enkelte BDS og ved Kreftregisteret. Opplysninger om screeningresultat, etterundersøkelser og patologiresultat kan hentes ut ved hvert BDS for kvalitetssikring av egen virksomhet. For nasjonal og lokal kvalitetssikring er denne

formen for selvevaluering viktig, da læringsaspektet knyttet til intervallkreft er viktig i et screeningprogram (1,2). Regranskningsstudier viser at omlag 20 % av undersøkelsene hadde tegn som kunne vært oppdaget ved screening, andelen varierte fra 1.3 til 35,9 % avhengig av studiedesign (29). For å stimulere læring bør det ved regranskning av mammogrammer være mulig å gjennomføre dette både med og uten informasjon om funn ved screening og diagnostiske bilder (fasiten).

Arbeid med brystdiagnostikk krever fokus på bildekvalitet, som ett av de viktigste aspektene (45,46). Det er etablerte krav til at radiologen må sørge for at krav til teknisk utstyr og arbeidsprosedyrer, til lys og lyd, og selvevaluering følges (1,2). Digitale mammogrammer fremstår som visuelt svært ulike, hva angår kontrast, metning av svart/hvitt, skarphet, støy og artefakter grunnet ulik bildeprosessering for tyder (45).

2.3.1 Kvalitetsmål for radiologisk virksomhet i Mammografiprogrammet

Et utdrag av kvalitetsmålene for radiologisk virksomhet er vist i tabellen under. Ønskede og akseptable mål i kvalitetsmanualen baseres på erfaringer fra prøveprosjekt i Mammografiprogrammet og den europeiske kvalitetsmanualen for organisert mammografiscreening (1,2). Begrepene i tabellen er beskrevet i *Kap2.2.1 Sensitivitet, spesifisitet, PPV*.

Tabell 3: Kvalitetsmål for radiologisk virksomhet i Mammografiprogrammet

Indikator	**Ønsket mål**	**Akseptabelt mål**
Mammografifunn som fører til at kvinnen selekteres til etterundersøkelse – første runde – påfølgende runder	 ≤5% 3%	 4-5% 3-4%
Deteksjonsrate – første runde – påfølgende runder Bakgrunnsinsidens – første runde – påfølgende runder	 7/1000 3/1000 >3 x insidens >1,5 x insidens	 6/1000 3/1000 3 x insidens 1,5 x insidens

PPV på bakgrunn av positiv mammografi – første runde – påfølgende runder	 ≥16% ≥16%	 >12% >12%
Andel Ductal carsinom in situ (DCIS)	10-20%	10%
Intervallkreft tilfeller pr 10 000 screenet	<16	<18

I hovedsak påpeker målene at det er ønskelig å etterinnkalle så få kvinner som mulig samtidig som en ønsker å oppdage krefttilfellene.

2.4 Etablering av kompetanse

Etablering av testsett for radiologer krever at vi har kjennskap til hvordan kompetanse etableres for yrkesgruppen. Kompetanse opparbeides gjennom det formelle utdanningssystemet ved undervisning og instruksjon, og videreutvikles i arbeidslivet og gjennom forskning (47).

2.4.1 Profesjonsstudiet og spesialistutdanning i medisin

Utdanning av helsepersonell i spesialisthelsetjenesten omfatter grunnutdanning, videre- og etterutdanning, turnustjeneste, spesialistutdanning og ulike typer kurs. Medisinstudiet er et 6-årig profesjonsstudium ved universitet og fører til graden cand. med. Studiet inneholder både teoretiske og praktiske fag. Praksisdelen av grunnutdanningen skjer i helseforetakene. Turnustjenesten er en del av grunnutdanningen og et vilkår for å få autorisasjon som lege (47). Statens autorisasjonskontor for helsepersonell (SAFH) gir profesjonsgodkjenning til helsepersonell (47). Innenfor medisin er det en rekke spesialistutdanninger. Legeforeningen har etablert et stort faglig apparat gjennom spesialitetskomiteer, spesialistrådet, kurskomiteer og utredningsutvalg for å sikre og utvikle kvalitet og fagutvikling av spesialister (48). Per i dag er spesialistutdanningen i stor grad styrt av ferdighetskrav, sjekklister, krav til operasjonslister osv. I tillegg er det målbeskrivelser for de ulike spesialistutdanningene. Dette er viktig for å beskrive kjerneinnhold i hver spesialitet, slik at den blir relativt lik uavhengig av utdanningssted.

2.4.2 Radiologi og diagnostikk

Radiologi er en av spesialistutdanningene innen medisin. Sentrale elementer i radiologisk metode er evnen til å oppdage patologiske forandringer (persepsjon) og tolkinger av disse (interpretasjon). Dette vil være subjektive ferdigheter. Erfaringsgrunnlag og faglig skjønn med blant annet avgrensning med tanke på normalvarianter og aldersbetingede forandringer er av en vesentlig betydning. Utfyllende kliniske opplysninger vil her kunne avgrense differensialdiagnostikken og være en viktig forutsetning for en mest mulig presis radiologisk diagnose. Premissene for bildediagnostikk og vurdering av undersøkelsenes egnethet ved ulike problemstillinger vil ofte være radiologens ansvar. Den diagnostiske prosess kan være komplisert da de bildemessige forandringene ofte er beskjedne og subtile, f.eks. usikre fortetninger på røntgen thorax og asymmetrier ved mammografi. Samarbeid mellom kliniker og radiolog er derfor helt nødvendig for å optimalisere den diagnostiske prosess og vil og være en viktig læringsarena for begge. Det tar omlag 12 år å erverve seg denne spesialiteten. En ytterligere spesialisering er kunnskap i mammadiagnostikk, som et fåtall av dagens radiologer besitter. Per i dag et det 77 radiologer som tyder screeningmammogrammer i Mammografiprogrammet. Det er ikke krav til å erverve seg kunnskap innen mammadiagnostikk i spesialistutdanningen for radiologi.

2.4.3 Læringsarenaer

Det meste av læring starter som oftest med at medisinstudenter observerer for så gradvis å prøve på egenhånd. Læring foregår også gjennom veiledning, som er den uformelle daglige læringen. For de ulike grupper vil dette være radiologisk arbeid, previsitter, visitter, morgenmøter, labratoriearbeid, poliklinisk arbeid og avdelingsarbeid. Her kan studentene, spørre og få umiddelbar tilbakemelding. Dette er ofte en god læringssituasjon. Dessuten foregår mye praktisk yrkesopplæring ved det vi omtaler som mester svenn, der en erfaren spesialist er mester og utdanningskandidaten er svenn. Dette egner seg godt for innøving av kliniske ferdigheter.

Læring innen radiologi etableres selvsagt etter hvert som en får erfaring og anvender sin kunnskap. Persepsjon og interpretasjon av røntgenbilder krever kunnskap og relevans for aktuell undersøkelse som kan være om sykdommens risikofaktorer, utbredelse og aldersvariasjoner. Observasjon krever også trening i form av at en må kjenne til ulike funn og normalvarianter av funn. Å analysere karakteristikker som form, mønster, hva som er patologi og det å kunne tolke bakgrunnsinformasjon krever medisinsk faglig kunnskap. Kommunikasjon med henvisende lege, også etter at bildene er beskrevet er viktig. Dette støtter nødvendigheten av læringsarenaer og viktigheten av både taus og eksplisitt kunnskap for å kunne øve inn ferdigheter (36,37). Mammaradiologisk læring etableres ved brystdiagnostisk senter, der det tverrfaglige arbeid er en viktig læringsarena.

Mammaradiologens læringskurve vil være påvirket av både antall utførte prosedyrer, tverrfaglig samarbeid med egen og andre yrkesgrupper og læringsmiljø ved egen arbeidsplass. Innen flere felt i spesialisthelsetjenesten kreves også et visst tyde- eller behandlingsvolum for å bli god på et fagfelt. En læringarena innen radiologisk virksomhet er bruk av kontrasignering med erfaren radiolog, der tilbakemelding ved uoverensstemmelser er viktig. I mammografiscreeningtyding kreves tverrfaglighet i arbeidet ved å se sammenheng med klinisk kontekst og dermed oppfølging og videre behandling. Dette krever god kommunikasjon med samarbeidende leger og er viktig med tanke på nivå og kontekst omkring kunnskapservervelse. Organisasjoner skaper kunnskap gjennom aksjon og interaksjon og dette er en dynamisk og til enhver tid pågående prosess (37). Kunnskapsskapelse foregår gjennom konvertering av taus og eksplisitt kunnskap og ved å etablere et godt miljø/bevissthet/team for kunnskapsutvikling. Ledende roller må med tydelighet formidle avdelingens/organisasjonens kunnskapsvisjon, slik at dette er forankret i organisasjonen. Formuleringen kan også komme fra eksternt hold, der gitte krav og kvalitetsparametre som innen mammografiscreening kan være en pådriver for måloppnåelse og kunnskapsskapelse. En formidling fra ”mellomledere” er at de som ”kunnskapsprodusenter” må aktivisere arenaer for kunnskapsutvikling. Etablering og

mulighet for å benytte et testsett ved et brystsenter understøttes av at utøvelse fremmes ved virtuell interaksjon med refleksjon gjennom handling via IT systemer.

2.4.4 Simulatorer og IT-systemer for ferdighetstrening

Innen flere områder av medisinen benyttes simulatorer og IT systemer for ferdighetstrening før oppstart og for erfarne brukere. Simulatortrening kan inngå som obligatorisk trening for enkelte spesialiteter, men ulike avdelinger og grupper vil ha forskjellige behov. Ferdighetstrening gynekologisk lapraskopisk kirurgi halverte operasjonstiden før reelle operasjoner (49). Ferdighetstrening kan gjøre manøvrering og håndtering av medisinske opplysninger lettere ved å benytte hypotetiske eller funksjonsbaserte elektroniske pasientjournaler (50). I Norge er mange grupper helsepersonell kjent med simuleringsprogram med Lærdals øvingsdukker for hjerte-/lungeredning (51). Simuleringsprogram kan etableres ved at treningen foregår i team eller at det trenes på kritiske medisinske situasjoner under veiledning av instruktør. Refleksjon og debrifing av egen og teamets innsats er et viktig læringsaspekt.

Etablering av kasuistikker i form av bildefunn ved egen avdeling eller for et fagforum innen radiologien kan være simulator for ferdighetstrening. Programvare og testsett med mammogrammer kan ha ulik læringsfokus som persepsjon og interpretasjon, med ulik sammensetning og andel av funn på mammogrammer, antall bilder, krav til registrerings og tilbakemeldingssystem.

2.5 Testsett av screeningmammogrammer som e-læring verktøy

Hvordan ferdigheter i tydeprosessen kan oppnås, opprettholdes eller forbedres er et nøkkelbegrep for radiologer og andre som tyder screeningmammogrammer. Et bidrag i dette kan være regelmessig egentesting ved å benytte programvare og testsett som et e-læringsverktøy for opplæring og kvalitetssikring– individuelt og for eget senter. Og bruk/egentesting av et slikt verktøy kan være et virkemiddel for bygging av kompetanse. Å øke den enkeltes kompetansenivå og ferdigheter er et mål for simulatortrening

Ved å fokusere på utfordringen ved å tyde screeningmammogramme har noen land utviklet gode rutiner og tilbakemeldingssystemer til radiologene for å kvalitetssikre screeningarbeidet. Enkelte land benytter også opplærings/testsett for at screeningtydere kan vurdere og forbedre egne ferdigheter. International Cancer Screening Network kartla juni 2012 medlemslandenes systemer for tilbakemelding til screeningtydere, der man blant annet spurte om deltagerlandenes bruk av testsett (52). Dette forankrer nytteverdien av programvare og testsett for ferdighetstrening og tilbakemeldingsystem.

I England benytters testsett for å måle radiologenes tydeferdigheter av screeningmammogrammer. Testsettene ”PERFORMS” (Personal Performance in Mammographic Screening) har vært benyttet siden 1991, et par år etter oppstart av det nasjonale screeningprogrammet ”The National Health Service Breast Screening Program (3,4,5,6,7). PERFORMS har i mange år fungert som et selvevalueringsverktøy. Screeningtydere gjennomgår et standard testsett sammensatt av ulikt utfordrende mammogrammer hvert annet år. De som benytter testsettene mottar tilbakemelding på egen tydingen umiddelbart etter test. Tilbakemeldingen gis i form av en grafisk fremstilling av ferdigheter sammenlignet med tydingen foretatt av et ekspertpanel av radiologer, en såkalt ”fasit”. Det er etablert testsett med ulik vanskelighetsgrad og malignitetssuspekthet i de bildemessige forandringer for bruk i opplæring, selvevaluering og for sertifisering. Dette er viktige aspekter både for å oppnå, opprettholde og forbedre kunnskap. I England har de også benyttet testsett med 1000 mammogrammer for å vurdere om radiografers tydeferdigheter kan sammenlignes med radiologer (53).

Testsett kan være sammensatt av screeningmammogrammer med ulike funn og med forskjellig registreringsløsning for klassifisering av diagnostiske funn. Fokus vil da være på persepsjon og interpretasjon av aktuelle funn. Testsett med screeningmammogrammer bør ideelt sett inneholde bilder med og uten

abnormaliteter og malignitetssuspekte funn. Ved bruk av testsett som simulator kan en kartlegge om det er spesielle lesjoner, kjerteltetthet, typer kalk eller forandringer som er spesielt vanskelig å tolke. Egen oppfølging og læring kan dermed iverksettes innenfor disse områdene – eller ved senere å sette sammen spesielt vanskelige testsett.

Kjerteltetthet er en selvstendig risikofaktor for brystkreft (54,55). For klassifisering av kjerteltetthet i brystet har American College of Radiology (ACR), som en del av Breast Imaging Reporting and Data System Atlas (BI-RADS® Atlas) utviklet et klassifiseringssystem med fire kategorier. Kategori 1 er sammensetningen i brystet nesten bare fettvev. Kategori 4 ansees å være ekstremt kjerteltette. Forenklet kan vi si at fettvev synes sort og kjertelvev hvitt på mammogrammer. Siden maligitetssuspekte funn oftest oppstår som hvite forteninger i et mammogram, vil persepsjon og klassifikasjon bli vanskeligere i et mammogram med tett kjertelvev. Tett kjertelvev kan redusere sensitiviteten på mammografityding fra 98 ved kjetreltetthet 1 til 48 % ved tetthet 4 (56).

Mammadiagnostisk rapportering og interpretasjon bør baseres på standardiserte og etablerte klassifikasjoner som BI-RADS (57,58,59,60), for å kunne kvalitetssikre og sammenligne resultater både nasjonalt og inetrnasjonalt (1,2,33,34,59). Ved retrospektivt å kategorisere lesjoner på mammogram med BI-RADS er interobservatørvariasjonen for PVV god og dermed en valid metode (59).

Godartede forteninger som cyster og firbroadenomer er vanligvis runde og velavgrensede. Ondartede fortetninger er vanligvis mer uregelmessige og med en mer uklar avgrensing samt innvekst i omkringliggende vev. Med evt innhold av uensartede mikroforkalkninger. Forkalkninger kan sees som som små spredte saltkorn – og dette kan være utrykk for mailignitet. Sannsynligheten for at at det er snakk om kreft økes når forkalkningene er grupperte og uensartede. De enkelte forkalkningers utseende

kan si noe om sannsynligheten for kreft. Ved at diagnostisk klassifisering må registrerer for masse og kalk gjøres samtidig en vurdering av lesjonens maligitetsuspekthet.

I en reell screeningsituasjon tydes omlag 60 undersøkelser per time, men tidsbruk varierer med radiologens tydeerfaring og erfaring med nye tydemodalitet (40). Har radiologen liten erfaring innen mammadiagnostikk øker tidsbruk ved tyding samt andelen falsk positive tyderersultater (61). En test/studiesetting vil medføre endret arbeidsmetode siden det kreves mer diagnostisk klassifisering. Det vil og være en forståelse av at andelen selekterte er høyere i et testsett enn i screeningsituasjon av en populasjon.

Det er interobservatørvariasjon i sensitivitet og spesifisitet i screeningtyding. Interobservatørvariasjon i BI-RADS terminologien er relativt liten (59). Ved å benytte testsett kan en òg måle om det er intraobservatørvariasjon, dersom samme testsett flere ganger av samme person.

3 Metode

Denne masteroppgaven består av flere praktiske deloppgaver. Deloppgavene er:

1. Kravutviklingsmetoder for programvaren med brukersentert design
2. Innhenting av tillatelser
3. Identifikasjon av et randomisert utvalg av screeningmammogrammer fra Mammografidatabasen
4. Innhente screeningmammogrammer fra fire sykehus PACS og sette sammen testsett
5. Utvikle et registreringssystem for tyding av mammogrammer
6. Brukbarhetstesting og regranskning
7. Utvikle et tilbakemeldingssystem for testsettet
8. Utprøving av testsettet

3.1 Kravutvikling av testsett og programvare

Systemutvikling kan gjøres med ulike tilnærmingsmetoder for å drive frem utviklingen av et system. Modellene gjenspeiler utviklingen av programvareutvikling. Det er flere hovedklasser av utviklingsmodeller hvor vannfall- og smidige metoder er to av disse (62). Vi har benyttet flere utviklingsmetoder i arbeidet.

Vi skisserte og planla mye av innhold i registreringsløsning, sammensetning av testsettene og fremdrift i starten av prosjektet. Vi etablerte en detaljert kravspesifikasjon, som en klassisk tilnærming etter vannfallsmetoden. Vannfallsmetoden beskrives som en planstyrt prosessmodell som gir prosjektledelsen god kontroll over utviklingsprosessens framdrift og kostnader (62). Metoden går ut på en grundig analyse og spesifisering av systemet før man starter med design og implementering av programvaren. Prosessmodellen deler utviklingsprosessen opp i tydelige faser som følger på hverandre og hver fase skal helst avsluttes og

dokumenteres før neste fase påbegynnes (62). Metodens svakhet er at den i liten grad legger opp til endringer i de avgjørelser som er tatt i forrige fase.

Realiteten i oppgaven og ønske om brukermedvirking i arbeidet viste at det var behov for å oppdatere og endre kravspesifikasjonen. Å utvikle et e-lærings testsett for en liten gruppe fagpersoner krever brukermedvirking (63). Mammaradiologene ble derfor involvert tidlig i prosessen. Utvikling av vårt program vil ikke bli optimalt uten brukerinvolvering, og kjennskap til at arbeidet ikke blir riktig første gang er viktig (63). Vi la opp til å få innspill på behov og sammensetning av bilder i testsett, og omfang av innhold i registreringsløsningen tidlig i prosessen. Metoden vi benyttet for utviklingsprosessen ble derfor basert på erfaring og medvirkning i en fleksibel utviklingsprosess.

3.1.1 Brukersentrert design

Vi benytter brukersentrert design for å fokusere på å tilfredsstille brukerens behov. Vi har fokusert på hvordan designet kan støtte bestemte brukere til å utføre gitte oppgaver i det å benytte et testsett, men vi har ikke fokusert på de teknologiske løsninger. Vi benytter metoden også med den hensikt å skape dialog med radiologene i utviklingsprosessen. Med tanke på brukskvalitet har vi fokusert på brukere i designprosessen, men radiologene har ikke deltatt i selve designet av brukergrensesnittet. Vi har involvert radiologene for å kartlegge domenet som brukergrensesnittet skal fungere i.
Vi har benyttet brukersentrert design for å etablere (63):

- Krav; der vi ønsker å forstå og spesifisere bruk av systemet
- Kravspesifikasjon: vi ønsket hjelp til å spesifisere bruker krav
- Design: vi ønsket hjelp til å etablere et system med hensiktsmessig design
- Evaluering: vi har gjennomført en brukerbasert vurdering av programmet

I utviklings og designprosessen vil vi vurdere hvordan vi kan få svar på problemstillingene. Innspill og ideer fra brukere er innhentet med kvantitativ og

kvalitativ måte. Den metodiske overveielse har bestått i å velge fremgangsmåten som passer best for den aktuelle problemstilling. Kvantitativ tilnærming er strukturert og systematisert, går i bredden og tar sikte på å formidle forklaringer. Ved bruk av kvantitativ metode kan informasjon formes til målbare enheter. Kvalitativ metode har til hensikt å fange opp mening og opplevelse som ikke lar seg tallfeste eller måle. Den kvalitative tilnærming går i dybden og har som formål å få frem sammenheng og helhet.

Siden vi har utviklet et nytt produkt, kreves "dybde" informasjon fra brukere. Slik informasjonen skaffes best gjennom kvalitative undersøkelser. Vi har benyttet flere metoder i utviklingsarbeidet, en metodetriangulering; vi har søkt løsninger for problemene fra forskjellige perspektiver. Vi har blandet kvalitativ og kvantitativ metode ved å benytte en form for fokusgruppeintervjuer og konkrete spørsmål om innhold, brukerkrav og brukbarhet.

Denne rapporten fungerer derfor som dokumentasjon for de etablerte testsettene og programvaren og kan dermed som grunnlag dersom systemet senere skal endres eller flyttes til en annen maskinvareplattform.

3.2 Brukermedvirkning og brukbarhetstesting

En av oppgavens problemstillinger er å vurdere om brukermedvirkning kan øke programmets anvendbarhet og brukertilfredshet. Som evalueringsmetode for å øke brukbarheten av systemet har vi benyttet brukbarhetstesting.

Første møte med radiologer var på Gardermoen mandag 25.april 2012. Vi presenterte vårt prosjekt for møtedeltagerene, som besto av 5 mammaradiologer fra de ulike helseregione.
Som verktøy har vi benyttet prototype, diskusjon og ustrukturerte intervjuer. Vår plan for masteroppgaven ble presentert med forslag til sammensetning av testsett og registreringsløsning, samt en ”operativ” versjon av programvaren tidlig i prosessen. Vi

gjennomførte da en brukertesting av programvaren (prototype) ved at noen radiologer kunne prøve en tidlig versjon av programmet. Dette var en horisontal prototype som viste totalsystemet uten særlig mye interaktivitet og funksjonalitet.

Vi spurte gruppen om holdninger og behov for testsett som opplærings-/e-lærings- og kvalitetssikringsverktøy. Vi ba om innspill til sammensetning av testsett, krav til uttrekk av undersøkelser fra mammadatabasen, sammensetning av mammogrammer i testsettene, funksjonelle brukerkrav og innspill til parametre i en registreringsløsning. Etter møtet analyserte vi tilbakemeldingene, arbeidet med kravspesifikasjonen og videre utvikling av programvaren.

3.3. Etikk og tillatelser

Vi etablerte en protokoll for arbeidet. Prosjektet ble 2.5.2012 meldt Kreftregisterets personvernombud som kvalitetssikringsprosjekt ”Melding av kvalitetssikringsprosjekt: Etablering av testsett for radiologer som tyder mammografibilder”. Begrunnelsen for etablering av testsett for radiologer handler om å finne ut og bidra til at " beste praksis følges". Det er radiologenes tydeferdigheter som skal måles. Vi vil benytte screeningundersøkelser kun fra gruppen kvinner som hadde samtykket til at lagring av personopplysninger ved normalt funn blir lagret permanent i Kreftregisteret.

Tillatelse til bruk av mammogrammer og planer for masteroppgaven ble 21.5.2012 tilrådt av Kreftregisterets personverombud ved Oslo Universitetssykehus før uttrekk. Prosjektet med å utvikle testsett er vurdert av personvernombudet til til å være kvalitetssikring av virksomheten i Mammografiprogrammet. Forutsetningen for tilrådingen er at bruk av bildene ikke skal identifisere enkeltpersoner, verken direkte eller indirekte.

For å kunne sette sammen testsettene som ønsket, er uttrekk av screeningundersøkelsene basert på individualisert screeninghistorikk, tyderesultater og kreftfunn og type apparat/leverandør. Bilder og resultater knyttet til kvinnens

screeningmammogrammer ble avidentifisert ved at bildene ble tildelt et løpenummer da bildene ble hentet ut fra PACS. Screeningundersøkelser er kun randomisert fra kvinner som har samtykket til lagring av personopplysninger. Krysslisten som kobler avidentifiserte data (bilder) slettes 31.12.2012. Kvinnens identitet er anonymisert når passordbeskyttet koblingsfil slettes.

Fire aktuelle Brystdiagnostiske senter (BDS) ble 23.5.2012 i forkant av randomiseringen forespurt om tillatelse til å hente bilder fra deres sykehus. Godkjenning fra alle fire BDS forelå før uttrekket ble gjort.

3.4 Identifikasjon av screeningmammogrammer til testsettene

Målgruppen i Mammografiprogrammet er kvinner i alderen 50-69 år. Alderskohortene benyttet i uttrekket er derfor fra årskullene 1941-1960. Uttrekket er gjort fra Mammografidatabasen på Kreftregisteret. Kvinnenes 11-sifrede personnummer ble brukt til å identifisere screeningundersøkelser som inngår i testsettene. Det ble trukket ut personopplysninger for kvinner med screeningmammogrammer for fire testsett, hvert sett var planlagt å skulle inneholde 100 screeningtester. Screeningundersøkelser skulle kun hentes blant digitale mammogrammer. De BDS hadde alle innstallert digitalt utstyr i 2009.

For å vurdere om ulik bildeprosessering blant utstyrsleverandørene og tyderes erfaring med bildekvalitet valgte vi å etablere fire testsett med mammogrammer fra kun en leverandør og ett screeningapparat i hvert testsett (GE, Siemens, Philips Sectra og Hologic).

Testsett kan ha en sammensetning som reflekterer en normal screening- og tydesituasjon med antatt friske kvinner i ulike aldersgrupper. I screeningsituasjon er det en fordeling av sanne positive (SP), sanne negative (SN), falske positive (FP) og falske negative (FN) screeningundersøkelser, *se kap 2.2.1 Sensitivitet, spesifisitet, PPV.*

Vi har valgt å trekke ut undersøkelser i gruppene SN og SP, med en aldersfordeling. Testsettene inneholder flere sanne positive enn i en screeningsituasjon, grunnet en lav andel krefttilfeller i screeningpopulasjonen. Reelt sanne positive i populasjonen ville gitt for få krefttilfeller i hvert testsett. Fordelingsnøkkelen mellom kategoriene ble fastsatt før randomiseringen. For å spesifisere dette uttrekket og krav til tyderesultater på testsettene var det nødvendig med screeninghistorikk på individnivå.

Vi inkluderer kun en undersøkelse av hver kvinne. Kun undersøkelser med unilateralt funn i henhold til tydescore ble inkludert. Bilder fra kvinner med tidligere brystkreft er ekskludert, da disse har forhøyet risiko for brystkreft i gjenværende bryst. Vi inkluderte ekstra undersøkelser for uttrekket for alle aldersgrupper og typer funn. Dette ble gjort for å kunne erstatte undersøkelser av kvinner med brystimplantat, kun ett bryst, at det var tatt flere enn fire bilder grunnet brystets størrelse etc. Identifikasjon og innhenting av mammogrammer var planlagt på bakgrunn av teoretisk og detaljert kunnskap om Mammografiprogrammet.

Vi har forutsatt at bildene som er trukket ut, er korrekt tydet og av tilfredsstillende kvalitet for det aktuelle formålet i masteroppgaven.

Vi inkluderte ikke falske negative og falske positive screeningundersøkelser i testsettene forde vi da måtte oppfølgende bilder for å kunne klassifisere disse korrekt. De fleste falske negative screeningundersøkelsene har sanne negative screeningmammogrammer. Bilder fra diagnosetidspunkt bør da med i settet for å ha en funksjon og læringsverdi i testsettet. Det samme er tilfellet for falske positive, der det ville vært hensiktsmessig å ha dagnostiske mammogrammer/videre utredning for korrekt klassifisering av tilfellet.

3.4.1 Sann negativ som test

Inklusjonskriterier for uttrekk av sanne negative: Undersøkelsene var tydet negativt (tydescore 1) av begge radiologene, på begge sider, i aktuell og forrige undersøkelse.

Et randomisert utvalg av negative screeningundersøkelser (*se Tabell 4*), pluss to undersøkelser, i hver aldersgruppe. Ønsket undersøkelsen var utført i perioden august-desember 2011. Vi bestemte at det skulle foreligger et negativt tyderesultat fra digitalt utstyr fra forrige screeningrunde, og de fire BDS var alle digitalisert i 2009. To ekstra undersøkelser trekkes ut i hver gruppe for å kunne erstatte bilder jfr. eksklusjonskriterier ovenfor.

3.4.2 Sann positiv som test

Inklusjonskriterier for uttrekk av sanne positive: Undersøkelsene var tydet positivt (tydescore 3 på samme side, 1 på den andre siden, av begge radiologene) og diagnosen brystkreft (enten type 2 som er DCIS eller type 3 som er infiltrerende) var påvist ved histologisk prøve. Det stilles ingen krav til screeningtesten før aktuell undersøkelse. Et randomisert utvalg av det antall sanne positive screeningundersøkelser pluss tre ekstra undersøkelser per aldersgruppe hentes ut (*se Tabell 4*). Den aktuelle undersøkelsen var utført i 2009 eller 2010, for å sikre at kreftfunn er ferdig kodet og registrert.

3.4.3 Sammensetning testsett

Under vises testsettenes ønskede sammensetning av sanne negative og sanne positive undersøkelser samlet og per testsett med aldersfordeling.

Tabell 4: Testsett sammensetning

OPTIMA	Sanne Negative ønsket	Ekstra sanne negative	Sanne Positive ønsket	Ekstra sanne positive
Testsett 1	80	8	20	12
Testsett 2	85	8	15	12
Testsett 3 2	85	8	15	12
Testsett 4	75	8	25	12
Totalt antall undersøkelser i fire testsett	325	8	75	
Ekstra undersøkelser i randomiseringen		*32*		*48*
Totalt antall undersøkelser	*357*		*123*	

Sammensetning og fordelingen i testsettene er planlagt som vist i *Tabellene 5 til 8*.

Tabell 5: Antall screeningundersøkelser etter tyderesultat og alder, Testsett 1 (GE)

Tyderesultat	**Totalt**	50-54 år	55-59 år	60-64 år	65-69 år
SN	80	22	21	16	21
SP	20	5	5	5	5
Undersøkelser	100	27	26	21	26

Tabell 6: Antall screeningundersøkelser etter tyderesultat og alder, Testsett 2 (Hologic)

Testsett 2	**Totalt**	50-54 år	55-59 år	60-64 år	65-69 år
SN	85	24	22	17	22
SP	15	6	3	3	3
Undersøkelser	100	30	25	20	25

Tabell 7: Antall screeningundersøkelser etter tyderesultat og alder, Testsett 3 (Philips Sectra)

Testsett 3	**Totalt**	50-54 år	55-59 år	60-64 år	65-69 år
SN	85	24	22	17	22
SP	15	6	3	3	3
Undersøkelser	100	30	25	20	25

Tabell 8: Antall screeningundersøkelser etter tyderesultat og alder, Testsett 4 (Siemens)

Testsett 4	**Totalt**	50-54 år	55-59 år	60-64 år	65-69 år
SN	75	22	21	16	16
SP	25	7	6	6	6
Undersøkelser	100	29	27	22	22

3.5 Innhenting av screeningmammogrammer

Screeningmammogrammer ble hentet ut fra fire ulike sykehus PACS. Avtaler, randomisering og fysisk uthenting ble gjennomført i perioden april – august 2012, der detaljene vises i *Tabell 9*.

Tabell 9: Fremdriftsplan for innhenting av screeningmammogrammer

Oppgave	Tid
Skriftlig avtale med gjeldende sykehus foreligger før uttrekk.	April - juni 2012
Etablere et løpenummer for alle screeningundersøkelsene og tyderesultat, med kobling til det 11-sifrede personnummeret.	Juni 2012
En liste med 11-sifrede personnummer bringes til fire ulike universitetssykehus hvor screeningmammogrammene hentes ut fra sykehusets PACS.	Juni – august 2012
Kvinnen og screeningmammogrammene identifiseres i PACS ved å benytte kvinnenes 11-sifrede personnummer samt invitasjonsnummer, som er unikt for det enkelte oppmøte.	Uthentings dato
Screeningmammogrammene anonymiseres ved fremhenting av bildene ved at kvinnenes identitet slettes og kun løpenummeret blir synlig på mammogrammene. Resultatet av screeningundersøkelsen gis samme løpenummer.	Uthentings dato
Mammogrammer til testsett Hologic (Vestfold) hentet vi ut selv ved BDS, Tønsberg fra SECTRA IDS7 PACS. Listen med personidentifikasjon på 119 kvinner og en ekstern hard disk medbragt.	27.6.2012
Haukeland tilbød seg å hente ut bildene for oss. Mammogrammer til testsett GE (Haukeland) ble hentet ut av en radiograf ved BDS, Haukeland fra deres Agfa PACS. Listen med personidentifikasjon på 119 kvinner, brukerveiledning og en ekstern hard disk sendt per post. Mammogrammene ble returnert på CDèr.	29.6.2012
Mammogrammer til testsett Siemens (Vestre Viken, Drammen) hentet vi ut selv ved BDS, Drammen fra Kodak PACS. Listen med personidentifikasjon på 119 kvinner og en ekstern hard disk medbragt.	27.7.2012
Mammogrammer til testsett Philips Sectra (Trøndelag St Olav) hentet vi ut selv ved BDS, Trondheim fra SECTRA IDS7 PACS. Listen med personidentifikasjon på 119 kvinner og en ekstern hard disk medbragt	17.8.2012
Koblingsnøkkelen mellom løpenummer og kvinnens ID slettes når bildesettene er ferdigstilt, jfr. rapportert dato til personvernombud.	31.12.2012

3.6 Utvikling av et registreringssystem

For at testsettet skal gi tyderne tilbakemelding på utført test må det registreres parametre som danner grunnlaget for tilbakemeldingen. Innholdet i registreringssystemet for tyding av mammogrammene er derfor viktig i utviklingen av systemet.

En av målsettingene var gi tyderne tilbakemelding på antall og andel av korrekte tydinger sammenlignet med ”fasit”, samt å gi tilbakemelding med ”fasit” for de tilfellene hvor tyders testresultat avviker fra ”fasit”. Vi ønsket i første omgang å etablere et registreringssystem som kunne brukes i en testsituasjon, men som også kan brukes ved en eventuell senere utvidelse, endring og distribusjon.

Vi ønsket å inkludere følgende parametre i registreringen:

1. Generell informasjon
    a. Identifikasjon av testsettet
    b. BrukerID
    c. Dato for tyding av settet
    d. Tidsbruk for tyding av hvert testsett
    e. Identifikasjon på om settet har vært lest av samme tyder tidligere
2. Mammografisk tetthet angis for begge bryst samlet
3. Mammografisk seleksjon: Negativ (tydescore 1)/Positiv (må etterundersøkes)
    a. Score (tydescore 2-5 for å angi malignitetssuspekthet)
    b. Positivt funn
        i. sidehenvisning
        ii. klassifisering av funn

Innholdet fra punkt 1 registreres automatisk ved pålogging. Punkt 2 og 3 registreres av radiolog ved gjennomføring av test.

3.7 Brukbarhetstesting

Brukbarhetstest er en samling teknikker for å optimalisere programmer i forhold til bruk, funksjonalitet og informasjonsgjenfinning. Kort fortalt handler dette om å gjøre navigasjon i programvaren og informasjonen som ligger der lettest mulig tilgjenglig for sluttbrukeren.

Tabell 10: Fremdriftsplan for brukerinvolvering

Involvering	Oppgave	Tid
Første involvering	Presentasjon med diskusjon innspill	25.4.2012
Andre involvering	Gjennomgang av brukerkrav og oppsettet for BI-RADS parametre i programvaren	4.10.2012
Tredje involvering	Regranskning av SP i tre testsett for å etablere ”fasit” for BI-RADS klassifisering, samt SUS test	16.10.2012
Test 1	Tyding av testsett 1 og SUS test	15.11.2012
Test 2	Tyding av testsett 2 og SUS test	16.11.2012
Test 3	Tyding av testsett 2 og SUS test	19.11.2012
Test 4	Tyding av testsett 2 og SUS test	29.11.2012

Andre møte med radiolog var på Ullevål Universitetssykehus torsdag 4.10.2012.
En meget erfaren mammaradiolog og leder Nasjonal Rådgivningsgruppe i Mammografiprogrammet deltok på møtet. Begge studenter og en veileder var på møtet.
Målsettingen for andre involvering var å regranske de sanne positive mammogrammene for å sette ”fasit” for klassifisering bildefunn. Regranskningen skulle gjennomføres ved å benytte testsett-programvare på medbrakt bærbar PC. Å vurdere mammogrammene på bærbar PC var ikke gjennomførbart, siden skjermen var av for dårlig oppløsning for persepsjon og klassifisering av funn. Vi forsøkte å koble opp programvaren mot Ullevåls 5K høyoppløsningstydeskjermer uten hell, grunnet manglende koblingsmulighet (ledning). Regranskning lot seg dermed ikke gjennomføre på det aktuelle tidspunkt.

Møtet skulle samtidig gjennomføre en brukertest av prototypen. Denne formen for brukbarhetstesting ble primært valgt for å kunne teste ut en design idé empirisk mot virkeligheten og sekundært for å kommunisere med brukere. Vi benyttet en såkalt High fidelity (Hi-fi) prototype med skjermbilder som ligner på sluttproduktet (64), dette er komplekse prototyper med mye detaljer. Slike benyttes ofte sent i prosjekter. Vi

benyttet en ”vertikal” prototype som går i dybden på en detalj dvs. å implementere nok interaktivitet og funksjonalitet til å kunne teste programvaren.

Møtet ble da benyttet til en gjennomgang av programvaren, brukerkrav fra radiolog og oppsettet for BI-RADS parametre i registreringsløsningen. Vi gjennomførte også en form for testing av skjermbildene. Metoden ble å stille spørsmål jfr *kap. 3.7.2. Observasjon og intervju* både i andre og tredje involvering. Innholdet i registreringsløsningen (BI-RADS parametrene) ble etter tilbakemelding fra radiolog omstrukturerert.

Tilbakemeldingene og resultatene medførte at vi omstrukturerte innholdet i registreringsløsningen, innførte muligheter for ytterligere zoom samt å kunne benytte window/level også under regranskning. Arbeidet ble sluttført før tredje involvering.

3.7.1 Brukskvalitet

Brukskvalitet defineres som er en kontekstavhengig egenskap av et produkt. Vi ønsker å måle brukskvaliteten av vårt produkt. Dette er en meningsfull metode siden det er en definert brukergruppe av systemet, hva testsettet skal benyttes til samt i hvilken sammenheng det skal brukes. Måling av brukskvalitet henger sammen med kartleggingen underveis i designprosessen. Et produkts brukskvalitet defineres ut fra tre faktorer som kan måles for bestemte brukere med bestemte mål i bestemte omgivelser (65). Disse faktorene er:

- Effektivitet: hvor godt fungerer produktet for bruker i dennes løsning av oppgaver på en tilfredsstillende og rask måte og sier noe om nøyaktighet og kompletthet. Det er effektivt ved at brukeren oppnår høy grad av produktivitet.
- Anvendbarhet: I hvor stor grad bidrar produktet til å løse de oppgaver bruker har mål om å utføre, eks brukere skal ikke bruke mer enn 10 min på oppgaven
- Subjektiv tilfredsstillelse: Tilfredshet er en subjektiv opplevelse av brukbarhet og sier noe om brukernes opplevelser, følelser og holdninger knyttet til produktet.

Tilfredshet kan også måles i positive/ negative kommentarer og hvor ofte brukerne gjør feil mens de bruker produktet.

Systemet har høy brukskvalitet når det er lett å lære, er effektivt, lett å huske, er feilfritt og tolerant samt behagelig og tilfredsstillende å bruke. Det finnes to typer brukskvalitet beregninger som kan fanges i løpet av en test. Disse beregningene inkluderer ytelsesdata (hva som faktisk skjedde) og preferanse data (hva deltakerne mente). Dette vil være nyttige innspill i videre prosess med utvikling av testsett.

3.7.2. Observasjon og intervju

Metoden for å kartlegge brukskvaliteten er å benytte teknikkene observasjon og intervju av radiolog.

Intervju er en samtale med konkret formål. Det benyttes ulike metodiske tilnærminger ved bruk av intervju avhengig av hva som skal undersøkes. Vi benytter delvis strukturerte intervjuer, som er det mest brukte innen kvalitativ forskning og kalles det kvalitative forskningsintervjuet. Relatert til oppgaven kan vi se på intervju som innsamlingsmetode for både kvalitative (tekst) og kvantitative (tall) data. Spørsmålene kan være både korte/konkrete og lange/omstendelige. I tillegg skal bruker passivt observere bruk av programvaren og stille spørsmål og gi oppgaver. Hensikten i intervjuer er som regel å avdekke ulike aktørers perspektiv på et fenomen og ved observasjon å avdekke ulike aspekter ved et fenomen. Ved å kombinere ulike kvalitative metoder som både intervjuer og observasjon er dette en variant av metodetriangulering.
Vi utarbeidet og benyttet en intervjuguide som var en tematisk rettledning med oversikt over viktige (åpne) spørsmål og et observasjonsskjema som skulle benyttes i møtet med radiologen. Guiden var vårt hjelpemiddel og ville sikre at den inneholder innledninger til spørsmål og fungere som manual for riktig utfylling. En intervjuguide skal være et hjelpemiddel for å stille de samme spørsmålene dersom flere tester programvaren.

Gjennomføring
Arbeidsstegene vi fulgte i andre og tredje radiologinvolvering var å følge Jacob Nilsens 10 punkt for gjennomføring av brukbarhetstesting (66).

- Vi introduserte oss selv
  - En student informerte om programvaren, og at det var programvare og innhold og ikke personen som skal testes.
- Testpersonen ble gitt konkrete oppgaver som skulle løses
- Tester ble oppfordret om å fortelle hva han gjorde og hvorfor ved å "tenke høyt"
- Testpersonen ble oppfordret til å være så ærlige som mulig: Det er slik han best kan hjelpe til med å forbedre programvaren.
- Begge studenter observerte og noterte både hva bruker sa og gjorde (om han/hun klikker på mye eller forstår programmet, hvor lang tid som brukes osv.)
- Etter testen spurte vi om designet virket hensiktsmessig og forståelig i bruk.

Oppgaver og observasjon
Vi vil vurdere de to brukerrelaterte målene anvendbarhet (hvor lett er det å tyde bilder med programvaren) og brukbarhet (subjektiv opplevelse) fra besvarelsene. Disse gis en score på en skala fra 1 til 7, hvor 1 angir lite tilfredshet (65). Spørsmålene ble formulert ut fra oppgaver som programvaren skulle løse og gjennomført i henhold til skjemaet under. Spørsmålene i *Tabell 11* benyttes som et bidrag i brukbarhetstesting og ble stilt ved andre involvering. Spørsmålene ble fremlagt og besvart av radiolog før vi avsluttet møtet.

Tabell 11: Anvendbarhet og Brukbarhet, del 1

Oppgave	Observasjon	Kommentar	Anvend-barhet	Bruk-barhet
Umiddelbart inntrykk av programvaren				
Synspunkter på parametrene i registreringsløsningen?				
Er skjermbildene brukervennlige?				
Er det noe i programmet du savner?				
Bør noe endres?				

3.8 Regranskning

Tredje møte med radiolog var på Ullevål Universitetssykehus tirsdag 16.10.2012. Av praktisk årsaker deltok kun samme radiolog som ved forrige (andre) involvering. Begge studenter og en veileder var på møtet.

Ny programvare ble koblet til Ullevåls 5K tydeskjermer. Radiologen var nå kjent med programvaren samt egne endringsønsker. Radiolog vurderte mammogrammene og satte fasit på tetthet og klassifisering av funn på sanne positive mammogrammer i testsettene i henhold til parametre i omstrukturert og omprogrammert programvare. Radiologen registrerte grad av malignitetssuspekthet (tydescore 2-5), samt klassifiserte funn etter modifisert BI-RADS i registreringsløsningen (57). En av studentene satt ved siden av radiolog og bekreftet topografi, histologisk morfologi og tumorutbredelse under klassifiseringen. I ett av settene var også veileder delaktig i klassifisering av diagnostiske parametre.

Spørsmålene ble fremlagt og besvart av radiolog umiddelbart etter bruk av programvaren. Spørsmålene i *Tabell 12* ble stilt ved tredje involvering. Se ellers *kap. 3.7.2. Observasjon og intervju.*

Tabell 12: Anvendbarhet og Brukbarhet, del 2.

Oppgave	Observasjon	Kommentar	Anvend-barhet	Bruk-barhet
Tror du at du vil kunne benytte programmet?				
Responstid mellom undersøkelsene				
Synspunkter på parametrene i registreringsløsningen?				
Er skjermbildene brukervennlige?				
Kan du gjennomføre en produktiv tyding?				

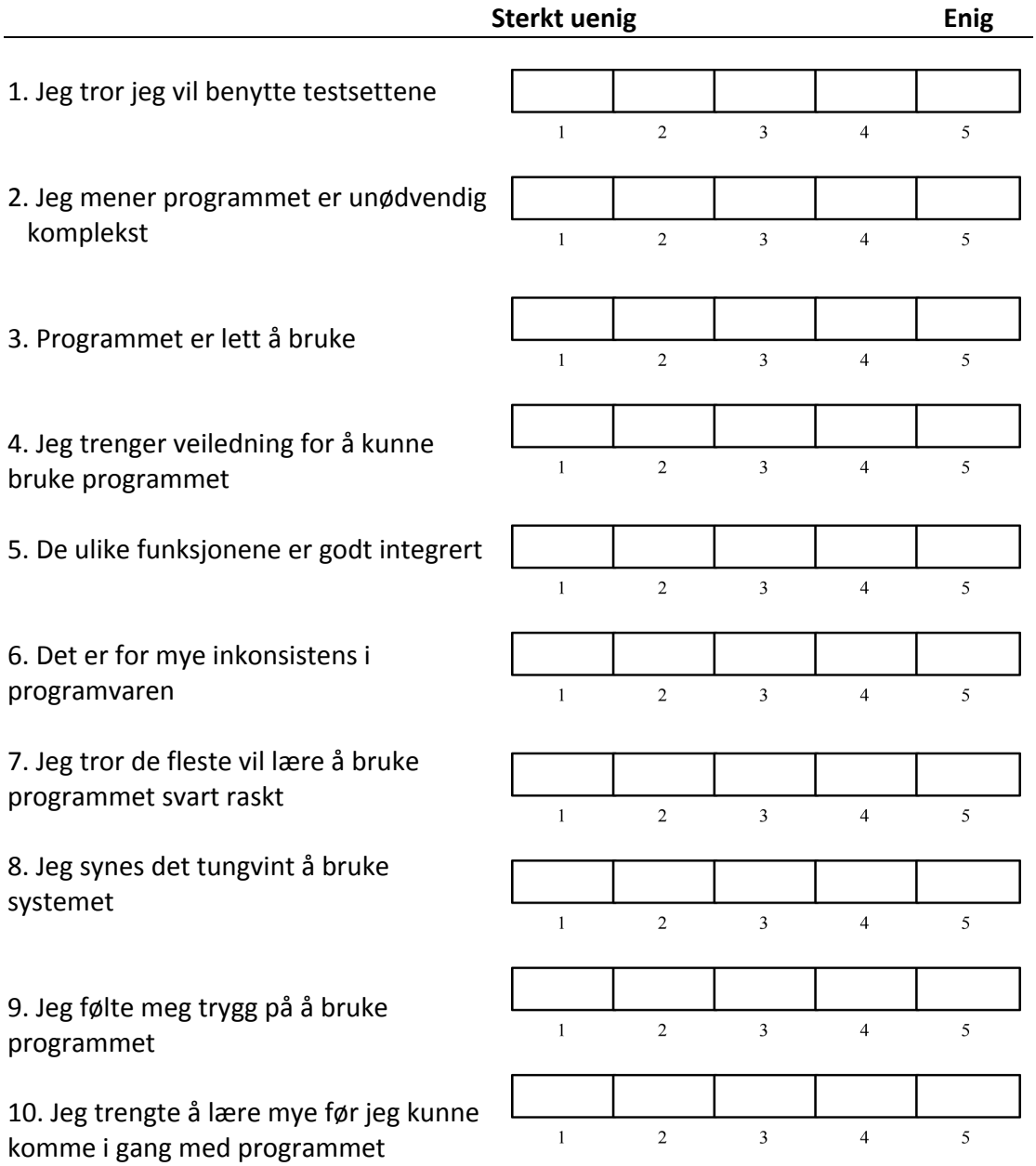
3.8.1 Brukervennlighetsskala

Kravene for å vurdere brukbarheten av systemer er en samling teknikker. Det kan bety at det hverken er kostnadseffektivt eller praktisk gjennomførbart. Vi ønsket en indikasjon på det generelle nivået av brukbarheten av systemet vi har utviklet. Vi valgte et tiltak som krever små anstrengelser og lave kostnader å samle inn og analyse av data. Vi ønsket å høre brukerens subjektive tilfredsstillelse. Den sier noe om brukerens subjektive opplevelse og vurdering av produktet etter å ha utført en brukbarhetstest.

For å få en oversikt over subjektiv brukervennlighet av programvaren benyttet vi ”System Usability Scale” (SUS) som er betraktet som en pålitelig og rimelig brukervennlighetsskala (67). SUS er en enkel, ti-punkt skala som gir en oversikt over subjektive vurderinger av brukervennlighet. Svarene angis på en Likert skala, som er en bipolar skala for å måle styrken av holdninger til gitte utsagn. Uttalelsene i skalaene er standardisert for denne aktuelle testen og respondentene angir grad av enighet eller uenighet med påstanden på en 5 punkt skala. SUS gir da et tall mellom 0 og 100 på om brukeren liker produktet etter å ha utført en brukbarhetstest.

Figuren under vises spørsmålene og brukervennlighetsskalaen. Radiologen besvarte de 10 spørsmålene i *Figur 1* umiddelbart etter regranskning av tre testsett.

Figur 1: "System Usability Scale" (SUS)



For å beregne SUS score summeres poengsummen, som vil variere fra 0-4 for hvert spørsmål. For poster 1,3,5,7 og 9 poengsummen bidraget er skalaen posisjon minus en. For poster 2,4,6,8 og 10, er bidraget 5 minus skalaen posisjon. Multipliser summen av

resultatet med 2,5 for å få den samlede verdien av SU. SUS score har en rekkevidde fra 0 til 100 poeng. Etter gjennomført test bør svarene analyserer og eventuelle endringer i programvaren foretas.

3.9 Tilbakemeldingssystem

Gjennomført tyding av et testsett resulterer i en rapport som kreeres på bakgrunn av tyderesultat og klassifisering som er gitt. Hensikten med tilbakemelding er å vise den enkelte tyders resultater etter angitte kriterier. Resultatene kreeres først når alle bildene i hvert testsett er ferdig tydet.

Rapporten skal gi umiddelbar tilbakemelding for antall korrekt selektert og antall korrekte klassifisering av positive funn. I tillegg skal inneholde rapporten opplysninger om når testsett er lest (dato og tid) og tidsbruk for tyding av testsett. Det er etablert mulighet for å gå tilbake og vurdere mammogrammer der eget resultat avviker fra fasit. På forespørsel kan resultater fra rapport skrives ut på papir eller eksporteres til andre applikasjoner som for eksempel Excel.

For masteroppgaven har vi valgt at alle data lagret på bl.a. ini-filer slik at behovet for database er fraværende.

3.10 Tyding av testsettet

Testsettene ble tydet av fire radiologer. Tre av radiologene er meget erfarne innen mammografiscreening og den siste er under opplæring En av de tre erfarne radiologene var involvert i første møte med radiologene, og ga sine synspunkter og innspill til utvikling av et testsett. De øvrige radiologene som testet ble forespurt av geografiske og praktiske årsaker.

Vår introduksjon til utprøving av testsettene vår innledningsvis å følge arbeidsstegene vi startet andre involvering med, *se kap 3.7.2. Observasjon og intervju.*

En radiolog tydet testsett 1 og tre radiologer tydet testsett 2.
Hver av radiologene mottok en tilbakemelding på egne resultater umiddelbart etter gjennomført test. SUS ble besvart på papir av radiologer som tydet testsett(67).

3.11 Utvikling av programvare

Til utvikling av programvare har vi valgt å skrive i C++ som er et høynivå, objektorientert programmeringsspråk som er utviklet fra programmeringsspråket C. C++ ble utviklet tidlig på 1980-tallet av dansken Bjarne Stroustrup og er den dag i dag et populært språk. Språket ble valgt grunnet studentenes tidligere erfaring med dette språket.

Som kompilator har vi brukt Borlands C++ Builder som er en robust og innholdsrik kompilator som inneholder de objektene som er nødvendig for å løse oppgaven. Kompilatoren ble valgt grunnet studentenes kunnskap og tidligere erfaring med denne samt kostnadsbesparelse.

4.0 Resultater

Hovedresultatet fra masteroppgaven er et system for tyding av screeningmammogrammer, som inkluderer både programvare og mammogrammer. Dette systemet har vi kalt OPTIMA. Etablering av OPTIMA er et kravutviklingsprosjekt som skal løse et behov for brukergruppen radiologer, nemlig å være et e-læringsverktøy for å tyde screeningmammogrammer. Vi har etablert detaljerte krav er viktig for å sikre at et program er riktig og i henhold til behov. Resultat fra vårt prosjekt er derfor samtlige kravspesifikasjoner for å hente ut testsettene og for utvikling av programvaren for tyding av mammogrammene. Kapittelet inneholder også resultater av innhenting av og regranskning av bilder. Spesifikasjonen er presentert som en prototype og med en liste av krav, både funksjonelle og enkelte ikke-funksjonelle. Resultatet er validert med bruk av brukervennlighetstest ved regranskning samt utprøving av testsett.

4.1 System spesifikasjon

Vi har valgt å utvikle en software som ikke krever noe installasjon hverken av databaser eller systemverktøy. Avhengigheter av «administrasjonsrettigheter» unngås derfor når programvaren taes i bruk. Systemet kan kjøres fra en ekstern harddisk på 1 terra byte eller kopieres over til en avdelingsintern partisjon. Det inneholder dog en ocx-fil som må integreres i lokalt operativsystem, som må være av typen Windows. Alle data er lagret på bl.a. ini-filer slik at behovet for database er fraværende.

Systemet er bygget opp med ulike deler. Disse er bruker, testsett med bilder og resultat og tilbakemeldinger.

- Brukerdelen er ment for å skille de ulike brukerne fra hverandre og inneholder data som navn, helsepersonell nummer (HPR-nummer) og et passord.
- Testsettet med bilder er for selve testen. Her ligger de fire ulike testsettene med referanser til de respektive hundre DICOM-bildene som hører til hvert av

testsettene. Programmet sørger for at bildene kommer opp i en tilfeldig rekkefølge. Under granskning av et sett kan man forstørre bilder, endre window/level og i tillegg er der et forstørrelsesglass. Man har også muligheten for å komme tilbake til utgangspunktet hvis man roter seg bort med window/level.

- Resultat og tilbakemeldinger til bruker etter gjennomført test. Etter test får man man umiddelbart opp alle svar som fraviker fra den oppsatte fasiten. Det er mulig å klikke på et case og få opp tilhørende bilder for å finne hva som skiller eget svar fra fasit. Resultatene kan hentes ut i en rapport som viser resultatene, rapporten kan lagres eller skrives ut.

4.1.1 Screeningmammogrammer i testsettene

Uthenting av screeningmammogrammer fra sykehus PACS foregikk i perioden juni til august 2012.

Det endelige resultatet av hva som foreligger av screeningmammogrammer i testsettene er resultatet av innhenting og gjennomgang av mammogrammer, samt regransking av de sanne positive mammogrammene.

Vi måtte eksludere de screeningundersøkelser som innehold bilateral brystkreft, til tross for at undersøkelsen var tyderegistrert med screeningfunn unilateralt, multifokalitet eller flere malignitetssuspekte lesjoner i ett bryst. Når en undersøkelse ble ekskludert, valgte vi neste løpenummer på listen til testsettet.

Tre av de fire testsettene ble regransket. Av praktiske årsaker ble ikke tid regranskning av testsett 4. Kun tre testsett kan benyttes til testing i masteroppgaven. En detaljert beskrivelse av antall og løpenummer inkludert og eksludert for alle testsettene ligger *Vedlegg 3 Screeningmammogrammer i testsettene.*

En oversikt over testsett etter utrekk, gjennomgang og regranskning vises i *Tabell 13.*

Tabell 13: Testsett sammensetning med eksludering

OPTIMA	Testsett 1 GE		Testsett 2 Hologic		Testsett 3 Sectra		Testsett 4 Siemens	
	SN	SP	SN	SP	SN	SP	SN	SP
Sammensetning, ønsket	80	20	85	15	85	15	75	25
Utrekk returnert KRG	87	32	93	27	93	27	83	37
Eksludert ved gjennomgang	13	5	5	2	1	1	8	1
Eksludert ved regranskning		1		8		3		
Sammensetning, endelig	**74**	**26**	**85**	**15**	**85**	**15**	**-**	**-**

Testsett 1
Testsett GE (Haukeland). Uttrekk av mammogrammene identifiserte 119 av 120 undersøkelser i PACS. Disse ble før brenning på CD èr avidentifisert deretter returnert Kreftregisteret. Hver CD ble gjennomgått av student. 18 undersøkelser ble ekskludert etter uttrekket blant kvinner med protese, pacemaker og mer enn fire mammogrammer fra screeningundersøkelsen. Alle bilder ble overført til harddisk. Ytterligere 1 undersøkelse ekskludert ved regranskning.

Etter gjennomgang hadde vi færre sanne negative undersøkelser en planlagt. Vi benytter derfor seks sanne positive mer enn planlagt i testsettet. Dette for å etablere et testsett med 100 screeningundersøkelser. Resultatene av fordeling i testsett 1 vises i *Vedlegg 3a Testsett GE (Haukeland).*

Testsett 2
Testsett Hologic (Vestfold) hentet vi ut selv ved BDS, Tønsberg. Listen med personidentifikasjon på 120 kvinner og en ekstern hard disk medbrakt. Alle screeningmammogrammene ble avidentifisert og medbrakt til Kreftregisteret.

Syv undersøkelser ble ekskludert etter uttrekket var blant kvinner med pacemaker og mer enn fire mammogrammer fra screeningundersøkelsen. Ytterligere 8 undersøkelser ble ekskludert ved regranskning, da det var vanskelig å sette fasit på

screeningmammogrammene grunnet multifokalitet og bilateral brystkreft (hvorav en ble oppdaget på etterundersøkelse). Resultatene 2 vises i *Vedlegg 3b Testsett 2 Hologic (Vestfold).*

Testsett 3
Testsett Philips Sectra (Trøndelag St Olav) hentet vi ut selv ved BDS, Trondheim. Listen med personidentifikasjon på 120 kvinner og en ekstern hard disk medbrakt. Dette uttrekket inneholdt falske positive undersøkelser. Første uttrekk og uthenting fra PACS ble derfor gjort på nytt av student. Uttrekk av mammogrammene identifiserte 120 undersøkelser som ble avidentifisert og medbrakt til Kreftregisteret.
To undersøkelser med protese og for mange mammogrammer ble ekskludert. Tre undersøkelser ble ekskludert ved regranskning. Resultatene vises i *Vedlegg 3c Testsett 3 Philips Sectra (Trøndelag St Olav).*

Testsett 4
Testsett Siemens (Vestre Viken, Drammen) hentet vi ut selv ved BDS, Drammen. Listen med personidentifikasjon på 120 kvinner og en ekstern hard disk medbrakt. Uttrekk av mammogrammene identifiserte 1 undersøkelser som ble avidentifisert og medbrakt til Kreftregisteret.
9 undersøkelser med protese, for mange mammogrammer og kvinne med ablatio ble ekskludert. Testsett 4 ble ikke regransket. Resultatene vises i *Vedlegg 3d Testsett 4 Siemens (Vestre Viken, Drammen).*

4.1.2 Use case og sekvensdiagram

Vi har benyttet use case modelleringsteknikk som gjerne brukes for å beskrive ønsket funksjonalitet i systemet. Use Case diagram viser hvordan systemet samarbeider med omgivelsene som f.eks brukere eller andre system (62). Systemer kan presenteres med flere use caser med ulike perspektiv, system eller brukere, oftest mange use case modeller for et system. Use casets styrke er å fremskaffe oversikt og forståelse for hvilke funksjonelle krav man ønsker å utvikle i systemet (68).

Sekvensdiagram beskriver hvordan flere objekter i et usecase samarbeider.

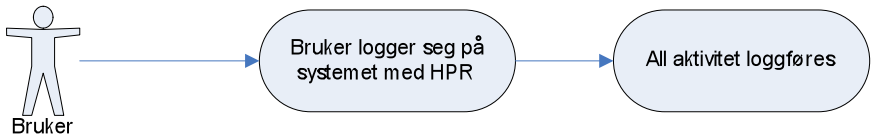
Diagrammene trenger ikke å vise hvordan selve samarbeidet utføres (68)

Sekvensdiagram er en mer abstrakt datamodell av scenariene og use casene [Fowler].

Sekvensdiagram kan også beskrive interaksjonene mellom brukere og objekter (62).

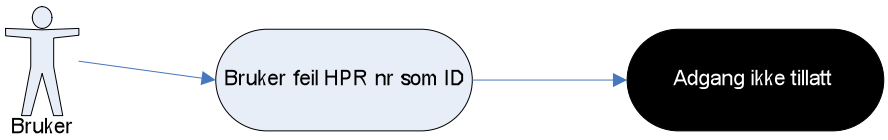
Vi har utarbeidet 4 use case og sekvensdiagram som vises i *Figur 2 til 5*.

Figur 2: Pålogging



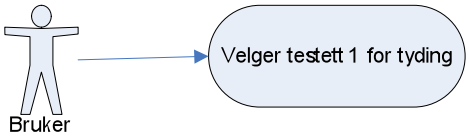
Use Case	**Pålogging**
Beskrivelse	Ny bruker logger seg på testsett
Aktører	Bruker, radiolog
Trigger	Ny bruker – godkjent HPR nr
Forventninger	Brukeren får tilgang til systemet
Forutsetninger	Systemeier har definert og implementert roller og foretatt nødvendig ajourhold på systemet.
Prosess	○ Brukeren logger på ○ Bruker kan utføre handlinger på system ○ All aktivitet loggføres
Avvik	Brukeren får ikke logget på systemet Bruker benytter ikke godkjent HPR
Resultat	Bruker har tilgang og kan tyde testsettet

Figur 3: Brukeridentifisering – misuse case



Use Case	Brukeridentifisering
Beskrivelse	Logger på uten korrekt HPR
Aktører	Bruker
Trigger	Feil HPR nr
Forventninger	Brukeren har ikke tilgang til systemet
Forutsetninger	HPR er definert og implementert for systemet. Bruker har ikke tjenestemessig lesetilgang
Prosess	○ Brukeren skal logge seg på programmet ○ Systemet ber bruker om å oppgi HPR ○ Behandleren skriver inn feil HPR nr ○ Tilgang til avvises
Avvik	Bruker oppgir ugyldig HPR og får tilgang
Resultat	Tilgang til programmet avvises

Figur 4: Tyding av testsett



Use Case	Brukeridentifisering
Beskrivelse	Ny bruker logger seg på testsett
Aktører	Bruker, radiolog
Trigger	Ny bruker – godkjent HPR nr
Forventninger	Brukeren får tilgang til systemet og kan velge ett eller flere testsett og gjennomføre tyding
Forutsetninger	Bruker har tjenestemessig tilgang
Prosess	○ Bruker åpner Testsett programmet ○ Velger hengeprotokoll ○ Velger testsett ○ Gjennomfører tyding ○ Mottar resultat for gjennomført tyding for ett sett

	eller totalt for alle
Avvik	Ikke mulig å velge hengeprotokoll Ikke mulig å tyde kun testsett 1
Resultat	Tilgang til programmet gis, tyding gjennomføres og tilbakemelding på gjennomført tyding

Figur 5: Rapporter

Use Case	Rapporter
Beskrivelse	Bruker henter ut informasjon fra rapportmodul og skriver det ut
Aktører	Bruker, radiolog som har tydet testsett
Trigger	Enten i form av predefinerte rapporter for hvert testsett eller for å hente ut bestemte opplysninger (eller kombinasjon av disse)
Forventninger	Bruker får hentet ut ønskede rapporter
Forutsetninger	Bruker har tilgang til data om seg selv
Prosess	○ Bruker logger på ○ Systemet godkjenner brukerID (HPR) ○ Systemet kontrollerer at brukers tilgang samsvarer med brukers bestilte rapportområde ○ Bruker angir hvilke opplysninger og/eller kombinasjoner av disse som skal presenteres ○ Informasjon skrives ut
Avvik	Bruker forsøker å hente ut rapporter fra en annen tyder han/hun ikke har tilgang til
Resultat	Mottar resultat på utskrift

4.1.3 Kravspesifikasjoner

Programvare opprettes for å løse et behov for en bruker. Vi har tilstrebet å etablere detaljerte krav for å sikre at programmet ble riktig og i henhold til behov.

Krav som gjelder for anvendelsesområdet - kan være funksjonelle og ikke funksjonelle. Vi har fokusert på å etablere funksjonelle krav, spesielt brukerkrav fra radiolog. Testsettets funksjonelle krav er konkrete krav som beskriver en ønsket tilstand og det programmet må utføre. Å etablere funksjonelle krav er det mest omfattende kapittelet i en kravspesifikasjon og her skal alle funksjonene i systemet beskrives. For å identifisere disse har vi vurdert brukers arbeid og arbeidsflyt, for å se hvilken del av arbeidet produktet best kan støtte. I tillegg er de beskrevet etter innspill fra radiologene ved første og andre involvering.

Brukerkrav til programmet

Vi har satt krav til systemet der det er kun aktørrolle, og spesifisere krav ut fra deres behov. Brukerkrav er uttrykt i et naturlig språk og ved å benytte diagrammer. Kravene kan beskrives ut fra systemets ønskede tjeneste og rammer. Vi hadde etablert en rekke krav til programmet, så kom radiologene med ytterligere kravene i første og andre radiologinvolvering. Enkelte krav definert ut fra krav til personvern, og tilråding fra personvernombud. Kravene er funksjonelle, dvs. hva systemet skal gjøre, og ikke-funksjonelle, dvs. egenskapene og omgivelsene til systemet.

Våre kravnummer er bygd opp slik: KR S1.

Forklaring på nummerering: K-krav. R-radiolog/O-Optima. S=Skal-krav / B=Bør; fortløpende løpenummer i kategori. Type F-funksjonell/IF-ikke-funkjonell.

Kolonnen ”Oppfyllt” i *Tabell 14* viser om kravet er oppfylt (Ja) eller ikke (Nei).

Tabell 14: Brukerkrav for OPTIMA

Kravnummer	Kravdefinisjon	Kravbeskrivelse	Type	Oppfyllt
KR S1	Personvern	Bildene skal være anonymisert	F	Ja
KR S2		Det skal kun være bilder av kvinner som har gitt sitt skriftlige samtykke til at personopplysninger knyttet til screeningundersøkelsen kan benyttes	F	Ja
KR S3	Testsett	Hvert testsett består av bilder fra en leverandør	F	Ja
KR S4		Det skal være en grei oversikt over de fire testsettene	F	Ja
KR S5		Hver undersøkelse skal inneholde cc og mlo bilateralt	F	Ja
KR S6		Hver undersøkelse skal kun inneholde ett sett på fire bilder per pasient	F	Ja
KR S7		Hver undersøkelse skal ha et løpenummer i testsettet	F	Ja
KR S8		Utvalget av screeningundersøkelser må bestå av bilder som sanne negative og sanne positive	F	Ja

KR S9		Utvalget av screeningundersøkelser må bestå av bilder fra ulike aldersgrupper	F	Ja
KR S10	View	Undersøkelsen skal åpnes med henge-protokoll fire bilder (mlo-mlo, cc-cc)	F	Ja
KR B1		Man må kunne velge mellom forskjellige henge-protokoller som 4-2-2, 4, 2-2	F	Nei
KR S11		Det skal være mulighet for zoom	F	Ja
KR S12		Det skal være mulighet for W/L	F	Ja
KR B2		Bildene kan vises på to skjermer	F	Nei
KR S13		Det skal være mulig å se hvor (hvilket nummer) man er i testsettet	F	Ja
KR S14		Rask respondering på bruker input	F	Ja
KR S15		Raskt skifte til neste billedsett på samme pasient	F	Ja
KR B3		Brukerveiledning må være tilgjengelig	F	Nei
KR B4		Tetthet registreres for alle undersøkelser	F	Ja
KR S16		Tyder angir seleksjon: malignitetssuspekthet og videre BIRAD klassifisering kreves	F	Ja
KR S17		Ved ferdig tydet undersøkelse må en hoppe automatisk til neste bildesett og løpenummer	F	Ja
KR S18		Ved ferdig tydet testsett, må neste testsett velges før start tyding av nytt sett	F	Ja
KR B5	Historikk	Historikk over hvilke testsett bruker tidligere har tydet	F	Ja

KR B6	Rapporter	Ønsker umiddelbar tilbakemelding på gjennomført tyding av testsett	F	Ja
KR B7		Være mulig å gå tilbake på undersøkelser som ikke er tydet ihht fasit	F	Ja

Systemkrav

Systemkrav er de kravene programvare stiller til maskinvaren i datamaskinen før programvare kan brukes og er ofte benyttet som en retningslinje istedenfor som et krav. Ofte gis to sett med systemkrav gitt: et minimumssett (som må være oppfylt før programvaren er brukbar på maskinvaren) og et sett med anbefalte systemkrav (for best ytelse). Datamaskiner med enda lavere spesifikasjoner enn minimumssettet kan noen ganger kjøre programvaren tilfredsstillende. Systemkravene for programvare har en tendens til å stige over tid. Systemkrav kan være vedlikehold og oppgradering av utviklingsverktøy, å etterkomme stilte sikkerhetskrav, å etablere en warm stand-by løsning, for å sikre kontinuerlig drift og minimal nedetid for systemet, sette krav til systemets infrastruktur som servere, nettverk, routere m.m., krav til lisenser omfatter rettighet til å bruke programvare til pc'er og servere.

Vi har beskrevet fem systemkrav for testsettene.

Tabell 15: Systemkrav for OPTIMA

Kravnummer	Kravdefinisjon	Kravbeskrivelse	Type	Oppfyllt
KO-S1	Tilgangskontroll	Nye brukere kan registrere seg selv	F	Ja
KO-S2		For å ha en unik nøkkel for hver bruker tas i bruk HPR	F	Ja
KO – B1	Oppslag	WEB-adresse for henting av HPR nummer er lett tilgjengelig	F	Nei
KO-S3	Operativsystem	OPTIMA må kunne kjøres Window XP eller nyere operativsystem	IF	Ja
KO-B2	Installasjon	Minimal installasjon nødvendig for å benytte systemet. Ingen bruk av database	IF	Ja

4.2 Validering

Validering av våre krav til programvaren skjedde ved å la brukere teste programvaren og gi en tilbakemelding ved hjelp av brukbarhetstesting. Resultater fra ”System Usability Scale” gjør at vi kan konkludere med at prototypen med gitte krav i kapittel *4.1* verdsettes av de som har testet OPTIMA.

4.2.1 Regranskning og brukbarhetstesting

Tilbakemeldinger og diskusjon med radiolog i andre involvering medførte endringer i programvarens oppsett av parametre, samt noe ny funksjonalitet. Programvarens skjermbilder og innholdet i registreringsløsningen før og etter omstrukturering er vist i *Vedlegg Vedlegg 4 Skjermbilderog Vedlegg 5 Parameterutvalg i registreringsløsning.*

Anvendbarhet (hvor lett er det å tyde bilder med programvaren) ble ikke besvart, siden dette ikke ble gjennomført. Svarene på brukbarehet ved subjektiv opplevelse varierte fra score 1 til 6 (*Tabell 16*). Svarene på anvendbarhet og brukbarehet fikk score 6 på alle spørsmål stillt i tredje involvering (*Tabell 16).*

Tabell 16: Spørsmål og svar fra andre involvering

Oppgave	Observasjon	Kommentar	Anvend-barhet	Bruk-barhet
Umiddelbart inntrykk av programvaren		Ser bra ut	*Ikke utført*	5
Synspunkter på parametrene i registreringsløsningen?	Ikke fornøyd	Må endres/ omstrukturer es		1
Er skjermbildene brukervennlige?	Prøver seg litt frem, men finner fort frem i programvaren	Ja. Lett å finne frem		6
Er det noe i programmet du savner?		Bedre zoom og enklere window/level		3
Bør noe endres?		Parametre for funn		3

I tredje involvering ble sanne positive screeningundersøkelser regransket ved å benytte eksisterende registreringsløsningen. Regranskning medførte ekskludering av ytterligere mammogrammer i testsettene, av årsaker som to malignitetssuspekte funn i ett bryst, bilateralt funn og mulitifokalitet. Detaljerte resultat fra er vist i kapittel *4.1.1 Screeningmammogrammer i testsettene* 4.1 og *Vedlegg Vedlegg 3 Screeningmammogrammer i testsettene*.

Tabell 17: Spørsmål og svar, tredje involvering

Oppgave	Observasjon	Kommentar	Anvend-barhet	Bruk-barhet
Tror du at du vil kunne benytte programmet		Ja	6	6
Responstid mellom undersøkelsene	Før ny undersøkelsene er klar til tyding	Gikk raskt i testsett 1 og 2, tregere i 3	6	6
Synspunkter på parametrene i registreringsløsningen?	Fjernet vi for mange parametre?	Fornøyd	6	6
Er skjermbildene brukervennlige?	Krever litt veiledning i oppstart	Ja	6	6
Kan du gjennomføre en produktiv tyding	Jobber effektivt – får bekreftet ”histologisk” fasit etter klassifisering	Ja, men ønsker mer hurtigtaster (mindre museklikk)	6	6

Resultatet fra radiologens angivelse av subjektiv brukervennlighet ved bruk av SUS ga en score på 92,5 %, *se Tabell 18*.

4.2.2 Utprøving av testsett med brukbarhetstesting

To testsett ble tydet. En radiolog tydet testsett 1 og tre radiologer tydet testsett 2. Testpersonene var kjent med at det var ulik leverandør til hvert testsett. Testsett en krevde justering av window/level for hver ny undersøkelse, selvom visningen gjenga orginalbilde. Testradiolog 1, 2 og 3 har lang erfaring med screeningtyding, testradiolog 4 er under opplæring. En av de erfarne radiologene hadde erfaring fra bruk av modifisert BI-RADS.

Vi ønsker av hensyn til testpersonene kun å angi hovedtrekk av funnene, disse presenteres i *Tabell 18*. Resultatene viser at det er en liten forskjell persepsjon av funn blant radiologene. Testene resulterte i en stor variasjon vises i antall selekterte. Mange

av de selekterte var definert som ”sanne negative” undersøkelser i testsettene, men ble av tester klassifisert som BI-RADS 2 konlusjon. Det var og variasjoner i angivelse av BI-RADS klassifisering av brystkrefttilfellene samholdt med fasit. Det er liten forskjell på resultatene blant de erfarne og den uerfarne radiologen.
Eksempel på variasjon i klassifisering blant radiologene var:
Fasit; Bløtdelspatologi – tumor – irregulær – spikulert – høy
Test: Bløtdelspatologi – tumor – irregulær – uskarp – Isodent

Resultat fra radiologers brukervennlighetstest både ved regranskning og utprøving av testsett ga en samlet SUS-score mellom 60 til 95 poeng. Detaljerte resultarer vises i flere tabeller i *Vedlegg 6 Brukervennlighetsskala*.

Tabell 18: Resultater fra tyding av testsett og brukbarhetstesting

		Testperson #			
	Regransker	***1***	***2***	***3***	***4***
Testsett	1,2,3	1	2	2	2
Brystkreft i testsett	-	20	15	15	15
Selektert totalt	-	28	36	21	24
Selektert SP	-	20	15	12	14
SUS score	**92,5**	**72,5**	**92,5**	**95,0**	**60,0**

Etter gjennomført tyding av et testsett mottok radiologen en tilbakemelding på resultat av egen tyding sammenlignet med fasit. Rapporten gir tilbakemelding for hver enkelt undersøkelse og resultatet av antall undersøkelser som er korrekt selektert som sanne positive og antall som er korrekt klassifisert i henhold til fasit. Rapporten inneholdt informasjon om tid benyttet for testen samt når rapporten er generert.

Rapportmodul inneholder en rullemeny, hvor man får opp en listevisning over de rapportene pålogget bruker i systemet har tilgang til. Etter at bruker har valgt ønsket rapport, får man mulighet til å velge ut resultat fra ulike testsett og om disse er gjennomført tidligere. Eksempel på rapport fra testsettene er lagret i *Vedlegg 4g*

Rapporter. Det er etablert hjelpetekst for radiologisk klassifisering av bildene med maligne funn. Egen vurdering kan sees i forhold til ”fasit” samtidig som screeningmammogrammene vises *Vedlegg 4h Se på sine feil.*

5.0 Diskusjon

Vår erfaring er at det var nødvendig med en bred teoretisk forankring, samt hensiktsmessig med kjennskap fagfeltet og arbeidsmetoder for å etablere ønsket system. Vi benyttet en fleksibel utviklingsprosess med brukerinvolvering, der vi fokuserte på å etablere gode brukerkrav for systemet. Diskusjonskapittelet er delt inn etter oppgavens problemstillinger.

5.1 Læringsteorier

Utvikling av et testsett med screeningmammogrammer krever at arbeidet forankres i flere erfaringsbaserte- og kunnskapsbaserte teorier. Det kreves kunnskap om screening, brystkreftepidemiologi, Mammografiprogrammet, metoder for evaluering av diagnostisk tester, hvordan kompetanse etableres for radiologer og bruk av e-læringssystemer. En viktig aspekt er kunnskap om betydningen av radiologers kompetanse og kvalitetsmål i forhold til mammografiscreening.

Å tyde screeningmammogrammer kontra diagnostisk mammografi krever ulike arbeidsmetoder. En screeningundersøkelse skal selektere hvem som krever ytterligere utredning. Diagnostisk utredning skal klarlegge en situasjon eller tilstand hos kvinnen. I våre testsett ønsket vi å fokusere både på evnen til å identifisere patologiske forandringer (persepsjon) og tolkinger av disse (interpretasjon). Å bli bedre i interpretasjon kan bidra til å øke kompetanse i å vurdere graden av malignitetssuspekthet på bildefunn, siden bildemessige forandringer ofte er beskjedne og subtile.

Det er etablerte retningslinjer, kvalitetsmål og krav til radiologisk virksomhet i screeningarbeid både på senter og individnivå i norsk og europeisk sammenheng (1,2). Korrekt persepsjonen er viktig for å oppnå kvalitetsmål som å detektere suspekte bildefunn/krefttilfeller (deteksjonsrate), PPV, sensitivitet og spesifisitet. Ved screening

i en generell populasjon er gjerne prevalensen av de sykdomsforandringer man leter etter lav, og mange kvinner med falske postivie funn selekteres til videre utredning (8). For å etablere testsett må en nødvendigvis ha en annen sammensetning enn i en normal screeningsituasjon. Dette kan være både en styrke og svakhet i settene. Antall funn i testsett var mye høyere enn i en reell screeningsituasjon for brystkreft. Bestemmende for sammensetning i uttrekk av screeningmammogrammer er av epidemiologisk karakter, bl.a. aldersjustert forekomst av brystkreft i Norge. Sammensetningen vi gjorde sikret at bryskreft i alle aktuelle aldergrupper var inkludert i testsettene.

Prediktiv verdi av en test avhenger av/varierer med forekomsten av aktuell sykdom i den befolkning pasienten kommer fra, og både PPV og deteksjonsrate øker med økende alder ved mammografiscreening. Mammografisk sensitivitet øker med alder, vesentlig grunnet redusert kjerteltetthet. Uttrekket inneholder i alle testsett flest undersøkelser fra den yngste kohorten, der det er flest falske positive screeningundersøkelser. I den yngste og prevalente screenede aldersgruppen har man ofte ikke gamle mammogrammer til sammenligning, noe som og er med på å gjøre tydingen av mammogrammene fra gruppen vanskeligere. Spesifikasjoner av uttrekkene ble gjort på bakgrunn av kjennskap til logistikk i Mammografiprogrammet og den sentral databasen, samt screeningkohortene og radiologenes arbeidsflyt. Samlet kunnskap forankret vårt uttrekk av screeningundersøkelser og sammensetning av testsettene. Testsettene må til enhver tid evalueres og etterstrebes å representere aktuell gruppe, så godt som mulig.

Norske og europeiske retningslinjer anbefaler at hver radiolog skal tyde minst 5000 screeningundersøkelser årlig for oppnå og opprettholde kompetanse (1,2). Volumet nås ikke av alle screeningtydere i Mammografiprogrammet per i dag (43). Den enkelte radiolog kan reelt sett tyde et høyere tydevolum, siden annen tydevirksomhet i offentlig eller privat regi ikke er medregnet. Uansett kan det å tyde et testsett bli et bidrag til å nå retningslinjer for anbefalt tydevolum. Radiologers erfaring og tydevolum

reduserer risikoen for en falsk positiv screeningundersøkelse (42,75), dette styrker tanken om at bruk av testsett kan være et bidrag til oppnå økt erfaring og tydevolum. Denne læringen kan også bidra til å redusere andelen falske positive screeningtester.

Radiologer er opptatt av å gi et mest mulig korrekt resultat, og hvordan de kan forbedre og opprettholde ferdigheter i å tyde mammogrammer. Kompetanse kan etableres ved tilpasset opplæring, ved å arbeide i gode tverrfaglige og etablerte miljøer, konferanser der fagfeller møtes og tverrfaglige faggrupper der resultater rapporteres og diskuteres. Å benytte et testsett kan være en arena og supplement for læring og kunnskapsformidling (37). Et testsett kan altså være et verktøy for egentrening og læring av mammografityding samt kvalitetssikring av virksomheten. Bruk av testsett med lagrede fasitsvar for diagnostisk klassifikasjon kan være et bidrag for radiologens evne til å treffe riktige konklusjoner. Tilbakemelding på gjennomført tyding gjør at en kan vurdere egne resultater. En engelsk studie viser at det er en signifikant positiv korrelasjon mellom det å benytte testsettet PERFORM med resultater fra et screeningprogram (69). Teorien støtter bruk av simulatorer for å oppnå læring og utvikline ferdigheter (49,47).

ICSN (Internasjonal Cancer Screening Network) har kartlagt hva som finnes av nasjonale og lokale programkrav til screeningprogram på verdensbasis (52). Sommeren 2012 ble det gjennomført en survey for å kartlegge type tilbakemeldinger som gis til radiologene. Spørreundersøkelsen kartla ønskede og akseptable mål for screeningtydere i startfasen og for mer erfarne tydere av screeningmammogrammer. De kartla hvilke type faglige anbefalinger som forelå for nye tydere, om det etter tyding av uerfarne benyttes skyggelesing eller gjennomgang av mentor, bruk av testsett, treningskurs eller annet. De kartla og om det forelå nasjonale eller lokale krav i det enkelte land; til tydevolum, antall, screening og diagnostiske undersøkelser og krav til oppnåelse av gitte kvalitetsindikatorer. Hvert land besvarte om den enkelte tyder gjennomgår mer trening dersom mål ikke nås. I USA benyttes begrepet ”fellowship training”, der læring foregår i etablerte fagmijøer/team. Det konluderes

med at denne metoden gir best deteksjonsrate og sensitivitet (70). Dette er i tråd med norske retningslinjer og viktighet av etablerte miljøer som læringsarenaer hvor egentesting ved bruk av e-læringsvertøy kan benyttes.

Å etablere testsett for tyding av mammografi kan ha ulike fokus. Vi mener det ikke finnes riktig eller feil sammensetning av testsett. Testsett kan etableres med fokus på opplæring eller egentesting, screening – eller klinisk tyding, persepsjon, klassifikasjon, eller e-læring. Ferdigheter kan testes jevnlig ved bruk av ulike typer testsett (3).

5.2 Kravspesifikasjoner

I utviklingsprosjekterer det viktig å ha kjennskap til ulike metoder for styring av prosjekter. For vårt prosjekt var det nødvendig å kunne benytte en smidig metode, der iterasjoner, samarbeid og tilpasning av prosesser gjennom hele prosjektet (62). Utviklingen og kravspesifikasjonene har bygget på kjennskap til fagfeltet og målsetting om å etablere et system som fungerer for bruker. Ved hjelp av et konstruktivt og iterativt samarbeid med bruker kunne vi håndtere endringer i programvaren underveis og etablere bedre kravspesifikasjoner. Involveringen var viktig for sikre god kvalitet og god brukeropplevelse av å benytte testsettene. Videre arbeid ble derfor preget av en iterativ utviklingsmetode, noe som passer godt for små prosjekter som dette (62). Utviklingsprosessen fulgte en evolusjonsmodell, der løsningen gradvis ble utviklet basert på erfaring og medvirkning i en mer fleksibel utviklingsprosess.

Det finnes mange måter å utvikle programvare på. Systemet er laget med tanke på minimal installasjon på lokal terminal grunnet forenklet bruk for radiologer uten bruk av avdelingens ikt-avdeling. Dette har fungert veldig bra, og har krevd kort tid fra vi ankom teststedet til testing kunne begynne.

Kravspesifikasjonens detaljnivå og innhold er avhengig av type prosjekt, hvem som utvikler/eier programvaren, krav til systemet samt økonomiske aspekter vil være av betydning. Vårt ønske har vært å tilrettelegge for en god arbeidsprosess hos bruker

(63). Vi har etablert en rekke funksjonelle brukerkrav til programmet, både for testsettene, programvare og rapporter. Kravspesifikasjonen vår fokuserer på å etablere gode brukerkrav som reflekterer hvordan radiologene arbeider når de tyder screeningmammogrammer. Enkelte brukerkrav er modellert ved hjelp av use case og sekvensdiagram, *se Figur 2 til Figur 5*. Vi har innfridd samtlige "Skal-krav" til systemet, men ikke alle "Bør-krav". Noen av "Bør-kravene" ble etterlyst av testerne og kunne med fordel vært innfridd for en bedre brukeropplevelse av programvaren.

Det utviklede systemet er et stand-alone system, hvor vi har etablert få ikke-funksjonelle krav og få systemkrav. Ikke funksjonelle krav er rammer for systemet og fokuserer på systemet egenskaper og mulighetene et produkt må ha, slik som ønsket utseende, brukskvalitet, pålitelighet og lover, regler, begrensninger og rammer. Kravene setter begrensninger for tjenestene produktet skal håndtere og kan være lite målbare forhold. Ikke-funksjonelle krav kan være inngangsterskel for bruk av systemet, læring: tid for å bruke systemet, behandling av feil og brukertilfredshet. Ønskede (målbare) kvaliteter på systemet (høy oppetid, svartid, feilprosent, antall samtidige brukere) kan ofte uttrykkes i tall (prosenter, antall, tid etc.).

Å ha godt kjennskap til brukergruppen, fagfeltet og databasen har vært svært nyttig for å kunne etablere uttrekk av screeningundersøkelser og sammensetning av testsettene. Det kreves en nøyaktig spesifikasjon for å hente ut ønsket sammensetning av screeningundersøkelsene. Dette var basert på tydescore fra hver radiolog, koblet mot sanne negative og histologisk verifisert sanne positive funn, både DCIS og infiltrerende tilfeller. Vi satte krav til funn i ett bryst med ulik malignitetsseuspekthet, ekskludering av tidligere brystkreft og undersøkelser med mer enn fire mammogrammer, valgt bildetakingsenhet på bakgrunn av benyttet digitalt utstyr og innføring av utstyret ved de enkelte bildetakingsenheter. For å kravspesifisere testsett med kun screeningundersøkelser er det nødvendig å ha kunnskap om screening som metode kontra diagnostisk utredning og kvalitetskrav for radiologisk virksomhet i mammografiscreening. For å kunne beskrive krav til uthenting fra databasen var det

nødvendig å kjenne logistikken i radiologers arbeidsmåter i Mammografiprogrammet og koding av kreft i mammografidatabasen. Det var også vært nødvendig å kjenne til samtykkekravet for lagring av negative opplysninger, både for korrekt utrekk samt i melding til Kreftregisterets personvernombud. For at bildene skulle være anonyme for radiolog var det viktig å spesifisere fremgangsmåte for avidentifisering ved uthenting, og senere anonymisering ved sletting av koblingsnøkkel. Mammogrammene er nå identifisert kun med løpenummer.

Testsettene består av mammogrammer fra en screeningundersøkelse. For screeningtyding er det anbefalt å inkludere foregående undersøkelse for sammenligning (1,44), slik som i en vanlig tydesituasjon. Testsettene inneholder kun sanne negative (SN) og sanne positive (SP) screeningfunn. Ved å etablere testsett som inneholder falske positive og falske negative mammogrammer, vil det være hensiktsmessig at radiologene ser gamle bilder til sammenligning. Spesielt dersom radiologen skal identifisere falske negative screeningundersøkelser som er oversett ved forrige undersøkelse. Man vet at de fleste intervallkrefttilfeller ikke var synlige på forrige screeningtest (29, 30). Vi mener det er et betydelig læringsaspekt med interpretasjon samt klassifisering av funn uten gamle bilder tilgjengelig. For å begrense mastergradsarbeidet og fokusere på de tekniske løsningene, valgte vi å inkludere kun SP og SN. Antall funn i testsett er som tidligere nevnt mye høyere enn i en reell screeningsituasjon. Eksempelvis ville da et testsett med 1000 bilder inneholdt omlag 5 brystkrefttilfeller.

Opplysninger som ønskes registrert for sanne positive funn er ikke registrert i samme omfang og måte i den sentrale mammografidatabasen som vi har utviklet i testsettets programvare. Å etablere konsensus ved å sette ”fasit” for bildefunn var derfor nødvendig for å kunne gi korrekt tilbakemelding på utført tyding og klassifisering av de ulike funn som benyttes testsettene. Etter gjennomført tyding av testsett vil tyders ferdigheter sammenlignes med ”fasit”, som danner grunnlaget for tilbakemeldingene i rapportmodul. Det styrker læringsaspektet å benytte standardiserte og etablerte

klassifikasjoner av funn (59), og det kan bidra til økt kunnskap om hvordan ulike funn har ulik grad av malignitetssuspekthet (57,58). Vårt valg av klassifiseringssystem er i tråd med Kreftregisteret fremtidige registrering. I løpet av det kommende år, vil de samme klassifiseringsparametre innføres på den samme måten for diagnostisk utredning for all brystkreft i Norge i Norsk Brystkreft Register.

Vi benyttet en prototype som et hjelpemiddel for å illustrere sentrale design-ideer for radiologene (64). Dette er en realistisk og rimelig metode og det er enkelt å gjøre modifikasjoner, samt at deltagelse etablerte en god dialog og rask tilbakemelding. Dette betyr at vi avklarte relevante problemer tidlig i utviklingsprosessen og at en prototype representerte en felles basis for kommunikasjon. En prototype vil kunne avdekke løst definerte eller uklare krav, hjelpe til med en raskere utvikling, evaluering og justering av programvaren. Det vil og kunne føre til utprøving av alternative tekniske løsninger. Vi benyttet første involvering som en form for fokusgruppeintervju, der vi spurte og diskuterte med potensielle brukere av systemet. Den uformelle formen på møtet åpnet for at deltakerne kunne komme med egne tema og innspill. En felles samtale kan belyse et bestemt emneområde, få frem brukernes erfaring og kunne gi mer konkret informasjon enn ved å intervjue ett og ett gruppemedlem eller ved å benytte spørreskjema. Det var dermed en nyttig form å benytte en prototype tidlig i prosessen for å få respons på egne ideer på utviklingen av testsett for radiologer og brukeres faktiske ønsker til systemet som skal utvikles.

I utviklingsprosjekter er det vanlig å utføre en risikoanalyse, og den viktigste er i forhold til pålitelighet. Pålitelighet kan deles i 4 hovedgrupper. Sikkerhet, tilgjengelighet, tilgjengelighet og stabilitet. Vi har vurdert risikofaktorer for utvikling av OPTIMA, men har valgt å ikke ta dette inn i resultatdelen av oppgaven. For et slikt stand-alone system vil det være lav sannsynlighet for hvor ofte en hendelse inntreffer og det vil ha ubetydelig konsekvens.

5.3 Praktiske forhold

For å kartlegge domenet måtte vi ha kjennskap til teori, hva finnes av testsett, hvordan fungerer disse, etablere mål og forbrede spørsmål til radiologer før vi startet designprosessen. Underveis i prosessen har vi opplevd at det vært en styrke for oss å ha god kjennskap og kommunikasjon med brukerne og BDS. Vi mener dette har lettet vårt arbeid og kommunikasjonen. Vi har blitt positivt mottatt ved forespørsel ved om tilgang til og uthenting av bilder, sammensetning av testsett samt brukerinvolvering og testing. Ved å kontakte radiologer tidlig i prosessen har vi fremskaffet en mer presis beskrivelse av oppgavene som skal løses og brukerne har hatt mulighet til å påvirke designet. Brukerens direkte involvering forventes å styrke deres aksept av testsettene. Ved involvering av andre må en påberegne ekstra tid for egen fremdriftsplan, siden oppgavene og skal passe med andre personer ”timeplan”. Vi ber om at ansatte utfører tilleggsoppgaver, og er derfor avhengig av velvillighet og frivillighet fra de involverte. Vi er tilfreds med at de praktiske omstendighetene for gjennomføring har ordnet seg på en god måte.

Innhenting av bilder en godkjenning fra de aktuelle sykehus. Før selve uthentingen var personalet ved BDS behjelpelige med å etablere kontakt med ”rette vedkommende” på radiologisk avdeling for veiledning for uttrekk fra PACS, ved tilrettelegging med en arbeidsstasjon en dag for student. Ved et sykehus var det nødvendig å registrere student som gjestebruker for å få tilgang til sykehus PACS. For to BDS var student selv kjent med PACS og hvordan bilder hentes frem og avidentifiseres fra PACS. Vi hentet ut mammogrammer ila sommeren, med redusert produksjon ved radiologisk avdeling. Etter at screeningmammogrammene var brakt ”hjem” krevdes en gjennomgang for å klargjøre undersøkelser som skulle inngå i testsett. Det var også nødvendig å regranske sanne positive funn, og vi ønsket å ferdigstille programvaren for dette formål. Dette gjorde det mulig å få en brukertest av systemet i andre involvering. En erfaren radiolog gjennomførte regranskningen og klassifiseringen av BI-RADS parametrene, at bildene ble vurdert sammen med histologisk verifisert kreftfunn styrker fasit på bildene.

En tydeprosess må gjennomføres på en effektiv måte, også ved bruk av testsett. Ved at det er lav responstid mellom undersøkelser og automatisk visning av neste undersøkelse kan radiolog arbeide effektivt. Dersom radiologene må vente på at bildene vises, vil fokus forsvinne fra tyding til irritasjon.
Mammogrammene må vises på høy-oppløslige skjermer, siden forandringer ofte er små og subtile. Slike skjermer benyttes ved brystdiagnostiske sentra og er kostbare. Å benytte mammogrammer med JPG bilder format i systemet ble diskutert og testet med radiologer, men konklusjonen var at bildene ikke ville bli av god nok kvalitet da spesielt ved forstørrelser.
Vi kjørte først et testsett fra laptop, men fant fort ut at dette ble alt for dårlig, både med hensyn til fart og til billedkvalitet. Vi fikk derfor kjørt både pre-test og de fire testsettene fra en stasjonær PACS-stasjon med hurtig maskinvare og høy-oppløslige skjermer. Det kreves orginalbilder og tydeskjermer med høy oppløsning for å oppnå høy sensitivitet i tydingen, også av testsett.

Arbeid med brystdiagnostikk krever fokus på bildekvalitet, som ett av de viktigste aspektene (1,2). Bildekvalitet er en "hjørnestein" i mammadiagnostikk, siden en leter etter små subtile forandringer. Kvalitetsstyring omfatter hele screening-kjeden, fra posisjoneringsteknikk, bildekvalitetskontroll og kvalitetskontroll av teknisk og fysiske forhold (45,46). Det er ulike leverandører av digital mammografi, noe som vil kunne påvirke valg av testsett. Det kreves en kritisk gjennomgang av mammogrammene, for å kvantifisere testsettene, samt kunne fungere som læringsverktøy for radiologene. Visuelt sett er testsettene noe ulike. Dette gjør at tester kan finne et testsett med billedkvaliteter man er kjent/ukjent med og gjøre testingen mer eller mindre utfordrende.

Siden vi har etablert et system laget med tanke på minimal installasjon på lokal terminal kan systemet kan kjøres fra en ekstern harddisk på 1 terra byte eller kopieres over til en avdelings-intern partisjon. Det inneholder en ocx-fil som må integreres i

lokalt operativsystem av typen Windows. Et scenario er at installasjon stoppes av dagens sikkerhetssystemer. Valget medfører at testsettet enkelt kan tilgjengeliggjøres for hele landet, og vil forenkle logistikken for radiolog med ønske om egentesting.

5.4 Brukermedvirkning

Et kravutviklingsprosjekt krever kunnskap og valg av metode(r) for å styre prosjektet. Kravutvikling og kravhåndtering for å etablere testsett omhandler det arbeidet som må utføres for å sikre at radiologenes behov for kvalitetskontroll og kompetanseheving blir omsatt til leveranse av et system som ivaretar dette behovet. Vår arbeid er basert på samarbeid og tilpasning av prosesser både etter planstyrt og smidig/iterativ metode (62). Vi har hatt et konstruktivt samarbeid med brukere av programvaren både for kravspesifisering og ved å benytte brukbarhetstesting.

Ved første involvering diskuterte vi behovet for og bruk av testsett for mammaradiologer. Radiologene både ønsket og så behovet for et norsk testsett. Ved å la brukere teste prototype tidlig i prosessen ga det brukerne en forståelse for tenkt produkt, og vi fikk avklart relevante problemer tidlig i utviklingsprosessen. Vår tidlig-versjonen av programvaren representerte en felles basis for kommunikasjon, og gjorde det enklere for brukerne å komme med innspill, krav og ønsker. Radiologene bidro med konkrete forslag til uttrekk og sammensetning av testsettene. Av forslag kan nevnes sammensetning testsett, innhold i registreringsløsning, hengeprotokoller, diagnostisk fasit og bruk av zoom. Det er nødvendig at sanne positive mammogrammer regranskes for korrekt fasit av diagnostikk parametre i henhold til modifisert BI-RADS klassifisering, siden positivt funn i testsettene krever denne klassifiseringsmetoden. Vår tanke om å etablere fire testsett ble positivt mottatt, kun med en påminnelse om å benytte mammogrammer fra ett og samme screeningapparat. Vi mener diskusjon og innspill viste at vi var på rett vei i utviklingsarbeidet. Dette ga videre inspirasjon til videre utvikling av systemet.

Radiologers påpekte at utvalg og sammensetning av brystkrefttilfeller burde inneholde enkelte DCIS tilfeller. Radiologene anbefalte å sette et krav til tydescore med usikker og høy grad av malignitetssuspekthet for DCIS og infiltrerende krefttilfeller settene. Dette forbedret vår spesifisering av uttrekk av screeningundersøkelser, siden diskusjonen bidro til beslutningen om å måtte angi malignitetssuspekthet på en score fra 2-5 for tilfeller som selekteres. For å gjøre egentestingen mer attraktiv for erfarne radiologer fulgte vi rådet om å blande inn brystkrefttilfeller selektert til etterundersøkelse med lav mistanke om malignitet. Disse kan være vanskeligere å detektere. Benevnes undersøkelsen som frisk selekteres den ikke, men gis tydescore 1.

De fleste av radiologene ønsket å ha gamle bilder fra fire år tidligere tilgjengelige i testsettene. Enkelte ønsket også muligheten for å ha kliniske opplysninger tilgjengelig, eksempelvis antegninger registrert av radiograf ved bildetaking. Dette kan være opplysninger om at kvinnen kjenner en kul, eller har en vorte, arr eller føflekk på brystet. Primært på grunn av prosjektets omfang tok vi en beslutning om å benytte en screeningundersøkelse uten angivelse av eventuelle funn, da det er tilstrekkelig i et system for opplæring og egentesting.

I andre og tredje involvering benyttet vi brukertesting som en evalueringsmetode der vi observerte og vurderte hvordan programvaren faktisk ble brukt av radiolog (65,66,67. Hensikten var å få en reell tilbakemelding fra bruker av programvaren. Vi mener det var viktig før testingen å formidle at vi ønsket hjelp til å teste ut et produkts brukervennlighet og ikke testpersonens ferdigheter (66). Vi formidlet også behovet for en faglig forankring og et friskt syn på dette, slik at vi kunne forbedre produktet. Vi observerte dermed raskt hva testpersonen brukte tid på og hva de utrykte skepsis til.

Vi brukte resultatene og tilbakemeldingene fra involveringene til å forbedre systemet. Fordelen med å benytte intervju som metode i involveringen var å innhente mer utfyllende data som ”faktainformasjon” og refleksjoner. Det bidro til å optimalisere innhold og brukervennlighet av programmet. I en brukbarhetstest vil man hovedsakelig

fremskaffe bekreftelse på om kravspesifikasjonene er tilfredsstillende eller om det er mangler som må tas med inn i implementeringsfasen. Brukbarhetstest i andre involvering ble utført ved å benytte en nesten ferdig versjon av programvaren (prototypen) og besto i inspeksjon av spesifikasjoner av brukerkrav og innhold i registreringsløsningen. Dette var den viktigste og mest reelle testen i løpet av prosjektperioden. Radiologenes medvirkning har hjulpet oss til å gjøre navigasjon i programvaren og informasjonen som ligger der lettest mulig tilgjenglig for sluttbrukeren. Siden vi i andre og tredje involvering benyttet en Hi-Fi prototype med mye detaljer fikk vi reelle tilbakemelding fra bruker (66). Testen var ikke avansert, og vi fikk raskt en indikasjon på hva som fungerer, og hva vi måtte gjøre noe med. Under andre involvering fikk vi helt konkrete tilbakemeldinger som omstrukturering av innholdet i registreringsløsning og ønske om forbedret zoom av bilder. Programvaren ble forbedret bl.a. med at perspektiv omkring de mest benyttede funn kunne stå øverst, for å lette registreringen ved funn. Dette medførte at vi endret innholdet i registreringsløsningen, samt ytterligere forberinger før neste involvering.

Vi mener at effektiviteten for bruk av programvaren har økt som følge av brukermedvirkningen i utviklingen testsettet. Å tyde testsettene kan etter involvering benyttes på en mer tilfredsstillende og rask måte, og arbeidet kan utføres med høy grad av produktivitet. Navigasjon i grensesnittet kunne vært ytterligere forenklet ved bruk av hurtigtaster. Å redusere bruk av PC-mus kan øke arbeidsflyten ytterligere. Programmet er anvendbart ved at radiologen kan løse oppgaven som er å tyde testsett av mammogrammer på en rask og effektiv måte. Testbruker utrykte tilfredsstillelse i sin holdning til testsettet. Både holdning og kommentarene var merkbart mer positiv ved tredje kontra andre involvering. Ved regranskningen lot det seg gjøre å benytte programvaren som et verktøy uten feil.
Radiologen som regransket har hatt sterk påvirkning på oppsettet av innholdet i registreringsløsningen. SUS ville trolig fått en helt annen score av radiologen dersom testen ble tatt ved andre involvering. Det ville vært en svakhet å benytte SUS skjema kun for en radiolog. SUS-skjema ble derfor benyttet ved egentesting i sluttfasen (67).

Dette styrker validiteten av resultatene fra tilbakemeldingene. Enhver bruker vil ha en subjektiv oppfatning av programmer, og ulike preferanser for bruk og innhold. Men, vi ser av brukertestingen at de involverte brukerne er godt fornøyd med systemet. Brukermedvirkningen har hjulpet oss med å utvikle og forbedre våre kravspesifikasjoner for sammensetning av testsettene, programvaren og innhold i registreringsløsningen. Vårt synspunkt er at utvikling og ferdigstilling hadde blitt vanskeligere og mindre hensiktsmessig dersom radiologene ikke hadde medvirket i kravspesifisering, brukertesting og reganskning. Gode brukerkrav og innspill gir et mer brukervennlig system. Vi kunne underveis i arbeidet satt av ytterligere tid til å diskutere konkrete brukerkrav med radiologene. I tillegg ville det med flere personer involvert kunne gitt oss flere tilbakemeldinger. Men, testpersonene er representative ved at de alle arbeider innenfor et smalt fagfelt og de ga ensartede tilbakemeldinger. Samtlige ”Skal” krav til programvaren er oppfylt. Ett av ”Bør” kravene, ”Bilder skal henge på to skjermer” er ikke oppfylt, noe som ble påpekt av flere testere. Som utvikler av en tjeneste blir en ofte ”blind” på svakheter ved tjenesten. Ved å sørge for brukermedvirkning bringers ofte nye perspektiv inn, som kan forbedre tjenesten.

5.5 Resultat av egentesting

Det var en liten forskjell i persepsjon blant de erfarne radiologene hva angikk å selektere krefttilfellene. Det er jo også en av grunnene til at dobbel uavhengig tyding benyttes som screeningpraksis for å øke deteksjonsraten (43). Det som varierte var antall seleksjoner. Radiologene var ikke kjent med antall krefttilfeller i testsettene, kun at settene var sammensatt av SP og SN screeningmammogrammer. Variasjon i seleksjon var trolig som en kunne forvente. Om dette hadde vært en reel screeningsituasjon ville et slikt resultat medført en rekke falske positive undersøkelser, vist i *Tabell 18*. Interobservatørvariasjon i mammografiscreening er kjent og i samsvar med resultatene *(*71,72,73*).* Det at mennesker vet at de er forsøkspersoner i et eksperiment har innvirkning på resultatet. Dette er kjent som Hawthorneeffekten, som dokumenterer at det å bli undersøkt, i seg selv gjør at personen har mer fokus.

Interobservatørvariasjon i BI-RADS terminologien er relativt liten (59). Testresultene viser en noe ujevn overenstsmmelse i BI-RADS klassifiseringen. Dette styrker teorien om behovet for at mer øvelse er hensiktsmessig siden denne klassifiseringen ikke er etablert praksis i Norge. Selv om norske radiologer er kjent med begrepsapparatet, er de allikevel uvant til med å benytte klassifikasjonsmetoden. For de selekterte med kreftfunn er det overensstemmelse mellom kalk og bløtdelspatologi/tumor, men overensstemmelsen mellom fasit på tumors form, avgrensning og tetthet samt kalkens morfologi og utbredelse er mindre overensstemmende. For BI-RADS klassifiseringen var det moderat overensstemmelse mellom form og margin av tumorfunn, noe dårligere overensstemmelse med distribusjon av kalk og sluttvurdering for angivelse av malignitetssuspekthet. Dårligst ut kom overensstemmelsen i angivelse av tetthet (høy, isodent, lav) på bløtdelspatologi. Resultatene er i tråd med gjennomgang av interobserver variasjonen innen BI-RADS terminologi, som viser at den er god og validerer amerikansk BI-RADS leksikon (59). Det er liten forskjell på resultatene blant de erfarne og den uerfarne radiologen, noe som tyder på at læringsaspektet er tilstede uansett erfaring. Ved å benytte testsett kan en òg måle om det er intraobservatørvariasjon, dersom samme testsett flere ganger av samme person.

Sammensetningen av screeningmammogrammer i testsettene er SP og SN, noe tetspersonene ble informert om før testen startet. Dette gjør at det ikke er anledning til å registrere BI-RADS funn for de som vi i uttrekket har tatt med som sanne negative og samtidig få en korrekt score. Sanne negative screeningundersøkelser i testsettene er gitt tydescore 1 av begge radiologer før uttrekk. Enkelte av disse inneholder eksempelvis kalk som testerene selekterte og klassifiserte som BI-RADS II. Dette forklarer trolig noe av det store antall seleksjoner, med diffus spredt beign kalk registrert med tydescore to av testerene.

For å oppnå høy sensitivitet og en lav andel falske positive, vil simulator som metode kunne være med på å optimalisere tyderes kunnskap ved å øve ferdigheter. Formål med testsett er å forbedre kvaliteten på screeningtyding lokalt samt å kunne måle

praksis mot etablerte standarder. Et testsett kan forsvares ut fra å benytte enkeltstående undersøkelser. En sammenligning med gamle mammogrammer ville sannsynligvis forbedret radiologens resultater, og ført til færre seleksjoner (44).

Alle benytter lengre tydetid enn i en reel screeningtyding. Dette er forventet siden det er en større andel krefttilfeller, som i tillegg skal klassifiseres på noe ukjent måte. ' Tydetid kan forøvrig være assosiert med økt falsk positive i mammografiscreening (61).

For testperson 1 og 2 var enkelte av mammogrammene feilmerket ved at tekst på front og skråbilde (cc og mlo) var byttet om. I tillegg overskygget DICOM teksten brystvev på enkelte bilder. Dette hadde ingen praktisk betydning for testen og er nå endret.

Styrker ved settene og testen

Screeningmammogrammene som viser brystkreft er regransket og funn på bildene har en fasit for parametre i registreringsløningen. Fasit er sett i sammenheng med angitt lokalisasjon, topografi samt histologisk verifisert morfologikode av patolog.

At det i registreringsløsningen i testsettene benyttes modifisert BI-RADS betyr at funn kan klassifiseres ihht et internasjonalt kodeverk, som benyttes i publikasjoner og litteratur. Denne klassifiseringsmetoden skal innføres som registreres som diagnostiske parametre i Norsk Brystkreft register, og skal taes i bruk i 2013. Å benytte flere og mer standardiserte parametre vil bety at testsettene kan fungere som et læringssett for mange norske mammaradiologer, både erfarne og uerfarne.

Kjerteltetthet i brystvev er en selvtendig risikofaktor for brystkreft (54,55), i tillegg til at det senker sentitiviteten av mammografi som metode (56). Det er lettere å overse brystkreft blant kvinner med høy kjerteltetthet, da dette kan maskere utseende på funn. Forandringer i brystvevet kan òg være vanskelige at tolke jfr. spesifisiteten. Å

registrere kjerteltetthet er dermed en viktig faktor, for læring av å detektere og klassifisere funn i bryster med tett kjertelvev.

Godartede forteninger, som cyster og firbroadenomer, er vanligvis runde og velavgrensede. Ondartede fortetninger er vanligvis mer uregelmessige og med en mer uklar avgrensing samt innvekst i omkringliggende vev med evt innhold av uensartede mikroforkalkninger. Forkalkninger kan sees som som små spredte saltkorn – og dette kan være utrykk for malignitet. Sannsynligheten for at at det er snakk om kreft økes når forkalkningene er grupperte og uensartede. De enkelte forkalkningers utseende kan òg si noe om sannsynligheten for kreft. Det er viktig å ha kunnskap om denne typen klassifisering av diagnostikken, som per d.d. ikke er lagret på samme måte i nasjonal database. Ved at diagnostisk klassifisering må registreres for fortetning og kalk, gjøres samtidig en vurdering av lesjonens maligitetsuspekthet. Funn av kalk på screeningmammogrammer representerer en utfordring i tolkningen av screeningmammografi (30). Det er viktig å gjennomgå mammografi tilhørende både screening og intervallkreft for å redusere oversette brystkrefttilfeller. En gjennomgang etter standardiserte klassifiseringsmetoder som BI-RADS klassifisering kan gjøre det mulig å vurdere resultater på tvers av fylker og screeningprogram.
En BI-RADS klassifisering krever at en ved funn av kalk må sier noe om distribusjon og morfologi. Klassifiseringene angir malignitetssuspekthet, det betyr å dignostisere mer enn kun å angi funn. Eksempelvis er diffus spredt kalk tilfeldig fordelt rundt i brystet, og type punktformig og amorfe er ofteste benign og bilateral. Den regionale kalken er spredt i et større område og ikke tilhørende en melkegang. Den er ofte benign dersom form ikke er av suspekte type. Lineær kalk ligger på linje, og er ofte suspekt på malignitet fordi dette antyder intraductal kalk. Segmental kalk er fordelt i et segment som tilsvarer en lobus og ofte antyder malignitet.

Vi mener denne vurderingen òg styrkes behovet for å starte med/lære seg radiologisk klassifisering av funn etter standardiserte metoder som i Optima.

Svakheter ved settene og testen

Mangel på mulighet til å registrere lokalisasjon av funn representerer trolig den største svakheten ved programvaren. Dette reduserer muligheten for å sikre at den som tyder testsettene klassifiserer riktig lesjon. Med kunnskap om hvor kreftsvulsten er lokalisert, kan man identifisere spesielle mønstre på mammogrammene som representerer en utfordring blant brystkrefttilfellene og andre funn.

Mangel på tidligere bilder kan være en svakhet. Inklusjon av gamle mammogrammer ville gjenskapt en mer reell tydesituasjon. Like fullt kan valget om en screeningundersøkelse forsvares fra et både et persepsjons, klassifikasjons og e-læringsperspektiv, og uansett vil prevalent screeningundersøkelse for kvinnen være uten tidligere mammogrammer til sammenligning.

En tredje svakhet er mammogrammene ble regransket/fasit er satt av kun en radiolog sammen med veileder. Veileder har en rekke publikasjoner på radilogisk diagnostikk og regranskning, men regranskning burde ideelt vært regransket i konsesnsus av to/tre radiologer.

Registreringsløsninge inneholder kun asymmetri med malign kalk, og er en svakhet i registreringsløsningen. Å kunne registrere kun asymmetri, distorsjon og asymmetri med kalk ble fjernet mellom andre og tredje involvering.

Radiologene rapporterte om uvant mye klikking/musebruk for å vurdere alle bildene, window/leve, zoom og bytte av hengeprotokoll. Ett av testsettene inneholder ikke tilfredsstillende window/level verdi, og radiologen må justere på hvert mammogram før tyding av bildene. Dette er tidkrevende. OPTIMA bør ha en funksjon for predefinering av window/level for undersøkelsen. Vi må vurdere om vi kan etablere mulighet for bruk av hurtigtaster, for å redusere bruk av PC-mus og klikking. Flere av radiolgene mente programmet er lett å bruke og at de trengte å lære lite før de kom i

gang med programmet. Men, flere har behov for en liten veiledning før testen starter. Vi bør lager derfor en liten brukerveiledning.

5.6 Videre planer

Programvare for testsettene er ferdigstilt, men ikke distribuert til alle landets BDS innenfor vår prosjektperiode. Vi må vurdere om testsettene skal distribueres eller om det skal etableres en fast test-lab hvor programvaren er koblet mot hensiktsmessige tydeskjermer i et egnet mørkt tyderom.

Vi har ikke innefor oppgavens rammer rukket å etablere fasit for det fjerde testsettet, ei heller å sette ”fasit” for tetthet for også de sanne negative screeningmammogrammene.

I Mammografiprogrammet gjennomføres fylkesbesøk. Representanter fra Nasjonal rådgivninggruppe og Kreftregisteret deltar på møtene sammen med ansatte ved BDS. Fylkets (BDS) resultater fra kvalitetsparametre blir presentert og diskutert. Gruppen kan komme med forslag til tiltak dersom enkeltparametre ikke tilfredsstiller kvalitetskravene eller retningslinjer ikke følges. Bruk av clinical audits kan sees på som en læringsarena for å forberde kvalitet (74) og fylkesbesøk kan vurderes benyttes til å benytte testsett for tyding.

Læring blant helsepersonell skjer ofte i grupper og ved veiledning. Jeg bør senere vurdere om IT systemet for testsett skal inneholde eller ta høyde for å etablere gruppefunksjonalitet. Dette kan gjenspeile en konsesussituasjon og være en god læringsarena (35,36,37).

Innenfor oppgaveperioden har vi ikke en ferdig plan for utskifting av bilder eller innhenting av nye, anonymiserte, ferdig tydere mammogrammer. Fremover i tid vil det være hensiktsmessig å etablere flere testsett for å kunne gi ytterligere basis for læring for brukere. Det vil da være naturlig å vurdere annen sammensetning av testsettene.

Radiologene påpekte muligheten for at de ulike BDS er bidragsytere med mammogrammer med spesielle funn og dermed mer e-læringfokus. Dette er i tråd med hvordan innhenting av bilder i utføres PERFORMS, England. Der etableres et nytt sett hvert år, dette distribueres for egentesting og ferdighetstrening. Dette vil medføre noe endring av programmets registreringsløsning og rapportmodul. Vi må etablere en plan for innhenting av anonymiserte bilder der fasit fra radiolog og patolog når bildene mottas. Dermed kan testsettene kunne benyttes av alle mammaradiologer, ikke bare de som arbeider i Mamografiprogrammet. Likefullt vil det være unødig å eteblere nye sett i nær fremtid, siden det vil ta tid før mange radiologer har tydet fire testsett.

Vi kan vurdere om vi skal etablere tilbakemelding til alle personer som har tydet testsettene innenfor en gitt tidsperiode skal få tilbakemelding på gjennomført tyding per sett sammenlignet med gjennomsnittet av alle tydere.

Vi må også vurdere om vi ønsker å etablere en funksjonalitet i programvaren, for at radiologen under tyding av bildene kan gå ett steg tilbake og se de forrige mammogrammene i samme løpenummer (undersøkelse) i hengeprotokollen. En mulighet er å etablere en tydesituasjon der gamle bilder (fire år tilbake) er tilgjengelig.

Det foreligger ikke sertifiseringskrav til den enkelte tyder i Mammografiprogrammet. Men, konsensus benyttes som en viktig læringsarena for å diskutere funn. En diskusjon for mammaradiologisk forening kan være om egentesting ved å benytte testesett kan være et bidrag dersom det er ønskelig med et akrediteringsprogram for tydere. Anbefalt tydevolum er vesentlig høyere i Europa enn i enn i USA. Studier diskuterer hvorvidt denne grensen, som er et ønsket mål i Europa, er for høyt (ref). Et etablert testsett kan bidra til en ny diskusjon om grensverdi for tydevolum.

Senere kan en teste et utvalg av personer med et visst tidsmellomrom for å analysere læringseffekt og intraobservatørvariasjon.

Et neste steg er å benytte mammogrammene med ny registreringsløsning for å vurdere radiograffaglig kvalitetsgradering av bildekvalitet. En registreringløsning for PGMI-klassifikasjon der radiograf klassifiserer bildekvalitet som Perfekte, Gode, Moderat gode og Inadekvate etter gitte parametre (1). Fokus her er i hovedsak kriterier som radiografen kan påvirke i screeningsituasjonen, og dette er arbeid som er etablert i produksjon. Men, et testsett kan være et bidrag for å lære seg klassifiseringen.

En annen mulighet er å benytte testsettene til subjektiv vurdering av bildekvalitet etter gitte parametre. Dette kan gjøre i tett samarbeid med Statens strålevern

6.0 Konklusjon

Mammografi har en sensitivitet på om lag 75% i screeningsammenheng. Sensitiviteten er i stor grad avhengig av radiologenes kompetanse. Teorien støtter bruk av testsett som simulatorer for å fremme radiologenes tydeferdigheter, samt for å etablere kompetanse innen klassifisering av mammografiske funn.

Vi har etablert OPTIMA, et system som består av en programvare for tyding av 100 screeningmammogrammer i fire ulike testsett. Hvert sett består av bilder fra en utstyrsleverandør. Testsettene er satt sammen av aldersfordelte sanne negative og sanne positive screeningundersøkelser. Spesifisering og sammensetning av utrekk krever opplysninger om screeninghistorikk, tyderesultater og type apparat/leverandør for hver enkelt screeningundersøkelse. Alle bilder ble avidentifisert ved uthenting fra sykehusenes bildearkiv (PACS). Sanne positive screeningmammogrammer ble gjennomgått av radiolog, som i tillegg til mammogrammene brukte histologiske opplysninger for å etablere ”fasit”. Breast Imaging-Reporting and Data System (BI-RADS) klassifisering ble brukt for å beskrive mammografiske funn. OPTIMA ble utviklet slik at det umiddelbart gir tilbakemelding på testers ferdigheter sammenlignet med ”fasit” når testen er gjennomført.

Programvaren krever ikke installasjon av databaser eller systemverktøy, eller «administrasjonsrettigheter». Pålogging kan finne sted fra en ekstern harddisk eller kopieres over til en avdelingsintern partisjon. For granskning av testsettene må det benyttes høyoppløslige tydeskjermer.

Programmet og testsettene er utviklet med brukerinvolvering. Brukermedvirkning har vært hensiktsmessig for å utvikle og forbedre våre kravspesifikasjoner for sammensetning av testsettene, programvaren og innhold i registreringsløsningen.

Gode brukerkrav og innspill har resultert i et brukervennlig system, noe radiologene bekrefter ved å gi programvaren høy score i brukervennlighetstesting.

Resultater fra fire radiologers tyding av ulike testsett viser liten variasjon i seleksjon av sanne positive funn. Derimot var det stor variasjon i antall selekterte, hvor mange av selekterte representerer negative funn. Det var stor interobservatørvariasjon i mammografisk klassifisering av sanne positive funn. Vi håper testsettene kan bidra til læring og heve kompetansen i forhold til BI-RADS klassifisering av mammografiske funn. Prosjektet kan i så måte betraktes som en del av kvalitetssikringen av det offenlige mammografiscreeningprogrammet i Norge.

7.0 Referanser

1. Mammografiprogrammet: Kvalitetsmanual. Kreftregisteret 2003. ISBN 82-90343-55-8.
2. Perry, N, Broeders, M, deWolf, C, Törnberg, S, Holland, R, and von Karsa, L. European guidelines for quality assurance in breast cancer screening and diagnosis. 2006. Belgium, European Communities, ISBN:92-79-01258-4.
3. PERFORMStm – PERsonal perFORmance in Mammographic Screening. *For Breast Screening Professionals* http://performs.lboro.ac.uk/for-breast-screening-professionals.htm (lesedato 18.2.2012)
4. Gale A.G. Maintaining quality in the UK breast screening program (keynote conference address) In D.J. Manning & C. Abbey (Eds.) SPIE Medical Imaging 2010: Image Perception, Observer Performance, and Technology Assessment (2010)
5. Gale A.G. & Scott H.: Measuring Radiology Performance in Breast Screening. In M.Michell (ed.) Contemporary Issues in Cancer Imaging – Breast Cancer, Cambridge UP, Cambridge, (2010)
6. Gale A.G.: PERFORMS – a self assessment scheme for radiologists in breast screening. In Seminars in Breast Disease: Improving and monitoring mammographic interpretative skills, 2003, 6(3), 148-152
7. Scott HJ, Gale AG. Breast screening: PERFORMS identifies key mammographic training needs.Br J Radiol. 2006 Dec;79 Spec No 2:S127-33.
8. Wilson JMG, Jungner G. *Principles and practice of screening for disease*. Geneva: WHO; 1968
9. Ekeberg O, Skjauff H, Karesen R. Screening for breast cancer is associated with a low degree of psychological distress. Breast 2001 Feb;10(1):20-4
10. Hafslund B, Nortvedt MW. Mammography screening from the perspective of quality of life: a review of the literature. Scand J Caring Sci 2009 Jan 7.
11. Hafslund B, Espehaug B, Nortvedt MW. Effects of False-Positive Results in a Breast Screening Program on Anxiety, Depression and Health-Related Quality of Life. Cancer Nurs. 2011 Nov 2.
12. Schou Bredal I, Kåresen R, Skaane P, Engelstad KS, Ekeberg O. Recall mammography and psychological distress. Eur J Cancer. 2012 Sep 27. pii: S0959-8049(12)00693-4. doi: 10.1016/j.ejca.2012.09.001.
13. Kalager M, Zelen M, Langmark F, Adami HO. Effect of screening mammography on breast-cancer mortality in Norway. N Engl J Med 2010 Sep 23;363(13):1203-10
14. Falk RS, Hofvind S, Skaane P. Overdiagnosis of Invasive Breast Cancer due to Mammography Screening. Ann Intern Med. 2012 Aug 7;157(3):219.
15. Hofvind S, Geller B, Vacek P, Thoresen S, Skaane P. Using the European guidelines to evaluate the Norwegian Breast Cancer Screening Program. Eur J Epidemiol. 2007;22(7):447-55.
16. Hofvind S, Sorum R, Thoresen S. Incidence and tumor characteristics of breast cancer diagnosed before and after implementation of a population-based screeningprogram. Acta Oncol. 2007 Sep 12:1-7

17. Hofvind S, Lee CI, Elmore JG. Stage-specific breast cancer incidence rates among participants and non-participants of a population-based mammographic screening program. Breast Cancer Res Treat. 2012 Aug;135(1):291-9. Epub 2012 Jul 26.
18. Olsen AH, Lynge E, Njor SH, Kumle M, Waaseth M, Braaten T, Lund E. Int J Cancer. Breast cancer mortality in Norway after the introduction of mammography screening. 2012 Apr 24.
19. International Agency for Research on Cancer *Breast Cancer Screening.* Vol. 7. Oxford: Oxford University Press; 2002. IARC Handbooks of Cancer Prevention
20. Broeders M. et al. The impact of mammographic screening on breast cancer mortality in Europe: a review of observational studies. *J Med Screen September 2012 19:14—25; doi:10.1258/jms.2012.012078*
21. Nyström L, Andersson I, Bjurstam N, Frisell J, Nordenskjold B, Rutqvist LE. Long-term effects of mammography screening: Updated overview of the Swedish randomised trials. *Lancer.* 2002;359:909–919.
22. The Swedish Organised Service Screening Evaluation Group Reduction in breast cancer mortality from organized service screening with mammography: 1. Further confirmation with extended data. *Cancer Epidemiol Biomarkers Prev.* 2006;15:45–51.
23. Gotsche PC, Nielsen M. Screening for breast cancer with mammography (review). Cochrane database of systematic reviews. *The Cochrane Library.* 2007. pp. 1–61.
24. Blanks RG, Moss SM, McGahan CE, Quinn MJ, Babb PJ. Effect of NHS breast screening programme on mortality from breast cancer in England and Wales, 1990-8: comparison of observed with predicted mortality. BMJ. 2000 Sep 16;321(7262):665-9.
25. Euroscreen Working Group. Summary of the evidence of breast cancer screening service outcomes in Europe and first estimate of the benefit and harm balance sheet. J Med Screen 2012;19(Suppl. 1):5–13.
26. St.prp nr.61 (1997-98) 5.2.1
Om nasjonal kreftplan og plan for utstyrsinvesteringer ved norske sykehus
27. Kreftregisteret. Nasjonale screeningprogram for kreft.
http://www.kreftregisteret.no. (lesedato 22.9.2011).
28. Hofvind S, Geller BM, Skelly J, Vacek PM. Sensitivity and specificity of mammographic screening as practiced in Vermont and Norway. Br J Radiol. 2012 Sep 19. [Epub ahead of print].
29. Hofvind S, Skaane P, Vitak B, Wang H, Thoresen S, Eriksen L, Bjørndal H, Braaten A, Bjurstam N. Influence of review design on percentages of missed interval breast cancers: retrospective study of interval cancers in a population-based screening program. Radiology. 2005 Nov;237(2):437-43.
30. Hoff SR, Samset JH, Abrahamsen AL, Vigeland E, Klepp O, Hofvind S. Missed and true interval and screen-detected breast cancers in a population based screening program Acad Radiol. 2011 Apr;18(4):454-60.
31. Thelle D. Innføring i epidemiologi 1998. Cappelen akademisk. ISBN 9788245600476
32. Cancer in Norway 2010 - Cancer incidence, mortality, survival and prevalence in Norway. Oslo: Cancer Registry of Norway, 2012

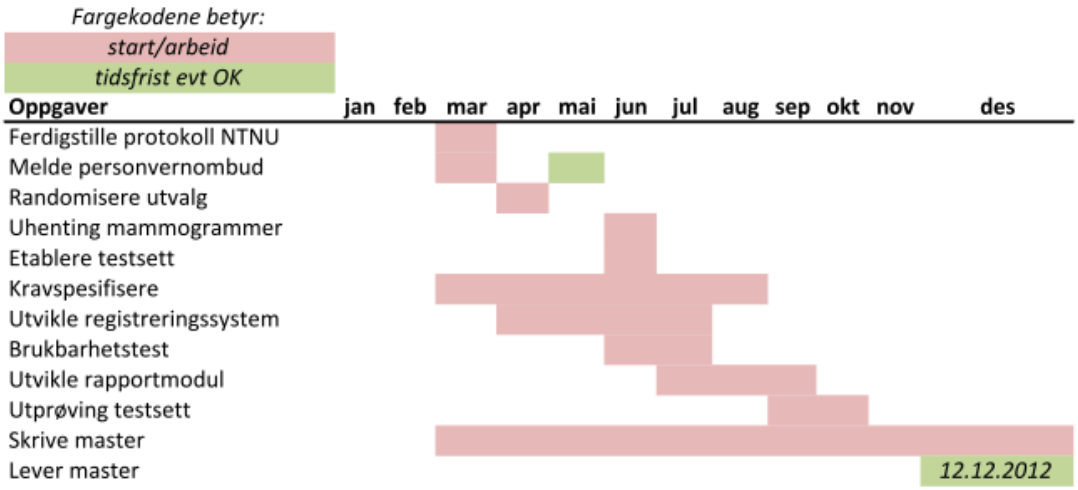
33. Trikalinos TA, Siebert U amd Lau J. Decision-Analytic Modeling to Evaluate Benefits and Harms of Medical Tests – White paper series. Rockville (MD): Agency for Healrhcare Research an Quality (US) 2009. (hhhp://www.nbci.nlm.nih.gov/books/NBK49464/)
34. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC; Standards for Reporting of Diagnostic Accuracy Group.Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. The Standards for Reporting of Diagnostic Accuracy Group.
35. Forrest CB. A typology of specialists' clinical roles. Arch Intern Med. 2009 Jun 8;169(11):1062-8.
36. Dreyfus SE (2004). The Five-Stage Model of Adult Skill Acquisition. Bulletin of Science, Technology Society 24(3):177-181
37. Nonaka I, Toyama R and Konno N. SECI, *Ba* and Leadership: a Unified Model of Dynamic Knowledge Creation. Long Range Planning 2000; vol 33 5-34.
38. Moss S.M et al. Is radiologists' volume of mammography reading related to accuracy? A critical review of the literature. Clin Radiol. 2005 Jun;60(6):623-6.
39. Miglioretti DL, Gard CC, Carney PA, Onega TL, Buist DS, Sickles EA, Kerlikowske K, Rosenberg RD, Yankaskas BC, Geller BM, Elmore JG. When radiologists perform best: the learning curve in screening mammogram interpretation. Radiology. 2009 Dec;253(3):632-40.
40. Skaane P, Hofvind S, Skjennald A. Randomized trial of screen-film versus full-field digital mammography with soft-copy reading in population-based screening program: follow-up and final results of Oslo II study. Radiology 2007 Sep;244(3):708-17.
41. Skaane P, Skjennald A, Young K, Egge E, Jebsen I, Sager EM, et al. Follow-up and final results of the Oslo I Study comparing screen-film mammography and full-field digital mammography with soft-copy reading. Acta Radiol 2005 Nov;46(7):679-89.
42. Alberdi RZ et al. Effect of radiologist experience on the risk of false-positive result in breast cancer screening programs. Eur Radiol (2011) 21:2083-2090.
43. Hofvind S, Geller BM, Rosenberg RD, Skaane P. Screening-detected breast cancers: discordant independent double reading in a population-based screening program. Radiology. 2009 Dec;253(3):652-60. Epub 2009 Sep 29.
44. Roelofs AA, Karssemeijer N, Wedekind N, Beck C, van Woudenberg S, Snoeren PR, Hendriks JH, Rosselli del Turco M, Bjurstam N, Junkermann H, Beijerinck D, Séradour B, Evertsz CJ Importance of comparison of current and prior mammograms in breast cancer screening. Radiology 2007 Jan;242(1):70-7.
45. Van Ongeval et al. Teaching syllabus for radiological aspects of breast cancer screening with igital Mammography. Radiat Prot Dosimetry 2008: 129(1-3):191-194.
46. Baert AL, Reiser MF, Hricak H and Knauth M (Eds).Medical Radiology –Diagnostic Imaging. Digital mammography ISBN 978-3-540-78449-4,2009.
47. Utdanning av helsepersonell. http://www.regjeringen.no/en/dep/hod/tema/sykehus/nokkeltall-og-fakta---

ny/sykehusenes-hovedoppgaver-/utdanning--av-helsepersonell.html?id=528641.lEST*) (lestedato 16.9.2011)*
48. Medisinstudiet. http://www.studentum.no/Medisin_46867.htm (lesedato 22.9.2011).
49. Larsen CR, Soerensen JL, Grantcharov TP, Dalsgaard T, Schouenborg L, Ottosen C, Schroeder TV, Ottesen BS. Effect of virtual reality training on laparoscopic surgery: randomised controlled trial. BMJ. 2009 May 14;338:b1802. doi: 10.1136/bmj.b1802.
50. Altman M. The clinical data repository: a challenge to medical student education. J Am Med Inform Assoc. 2007 Nov-Dec;14(6):697-9. Epub 2007 Aug 21.
51. Laerdal – helping save lifes http://www.laerdal.com/no/nav/336/Kontakter (lesedato 25/10/2011)
52. Kartlegging av ICSN`s medlemslands tilbakemeldingssystemer til screeningtydere http://appliedresearch.cancer.gov/icsn/projects/audit.html. Lesedato 5. September 2012
53. Wivell, G et al (2003). Can Radiographers Read Screening Mammograms? Clinical Radiology 58, 63—67)
54. Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, Jong RA, Hislop G, Chiarelli A, Minkin S, Yaffe MJ Mammographic density and the risk and detection of breast cancer. N Engl J Med. 2007 Jan 18;356(3):227-36.
55. Ellingjord-Dale M, Lee E, Couto E, Ozhand A, Qureshi SA, Hofvind S, Van Den Berg DJ, Akslen LA, Grotmol T, Ursin G. Polymorphisms in hormone metabolism and growth factor genes and mammographic density in Norwegian postmenopausal hormone therapy users and non-users. Breast Cancer Res. 2012 Oct 24;14(5):R135. [Epub ahead of print].
56. Kolb TM, Lichy J, Newhouse JH. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. Radiology. 2002 Oct;225(1):165-75.
57. D`Orsi CJ, Basset LW, Berg WA et al. BI-RADS Mammography, 4th edition in:
58. D`Orsi CJ, Mendelson EB, Ikeda DM et al. Breast Imaging Reporting and Data System: ACR BI-RADS – Breast Imaging Atlas, Reston, VA, American College of Radiology, 2003
59. Lazarus E et al. BI-RADS Lexicon for US and Mammography: Interobserver Variability and Positive Predictive Value. Radiology 2006 may, 239, 385-391
60. Taplin SH. et al. Concordance of Breast Imaging Reporting and Data System Assessments and Management Recommendations in Screening MammographyRadiology 2002 feb, 222, 529-535.
61. Carney PA, Bogart TA, Geller BM, Haneuse S, Kerlikowske K, Buist DS, Smith R, Rosenberg R, Yankaskas BC, Onega T, Miglioretti DLAJR Am J Roentgenol. Association between time spent interpreting, level of confidence, and accuracy of screening mammography. 2012 Apr;198(4):970-8.
62. Ian Sommerville. Software Engineering, 9th edt. Pearson ISBN 10: 0-13-705346-0
63. ISO 13407 Human centred design processes for interactive systems

64. Preece et al, 1994. Human-Computer Interaction. Wokingham, UK: Addison-Wesley. ISBN 0-201-62769-8
65. ISO 9241-11:1998 Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability
66. Nielsen J et al, 1994. Heuristic evaluation, *Usability Inspection Methods*, John Wiley & Sons, New York,NY.
67. Brukervennlighstesskala SUS http://en.wikipedia.org/wiki/System_Usability_Scale, Lesedato 5. September 2012
68. Fowler M, 2003. UML Distilled. A brief guide to the standard object modeling language. 3rd etd. ISBN-10: 0321193687 | ISBN-13: 978-0321193681
69. Scott H.J et al. The relationship between real life breast screening and an annual self assessment scheme. IN: Medical Imaging 2009: Image Perception, Observer Performance, and Technology Assessment, edited by Berkman Sahiner and David J. Manning, Proc. SPIE 7263,72631E.
70. Elmore JG, Jackson SL, Abraham L, Miglioretti DL, Carney PA, Geller BM, Yankaskas BC, Kerlikowske K, Onega T, Rosenberg RD, Sickles EA, Buist DS. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy.). Radiology. 2009 Dec;253(3):641-51. Epub 2009 Oct 28.
71. Skaane P, Young K, Skjennald A. Population-based mammography screening: comparison of screen-film and full-field digital mammography with soft-copy reading--Oslo I study. Radiology. 2003 Dec;229(3):877-84. Epub 2003 Oct 23.
72. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. *Variability in radiologists' interpretation of mammograms. N Engl J Med 1994*; *331*:*1493*-1499.
73. Beam CA, Layde PM, Sullivan DC. *Variability in the interpretation of screening mammograms by US radiologists. Arch Intern Med 1996*; *156*:*209*-213.
74. Flottorp et al. Using audit and feedback to health professionals to improve the quality and safety of health care. WHO Health Evidence Network, 2010.
75. Buist DS, Anderson ML, Haneuse SJ, Sickles EA, Smith RA, Carney PA, Taplin SH, Rosenberg RD, Geller BM, Onega TL, Monsees BS, Bassett LW, Yankaskas BC, Elmore JG, Kerlikowske K, Miglioretti DL. Influence of annual interpretive volume on screening mammography performance in the United States. Radiology. 2011 Apr 259(1):72-84.

Vedlegg

Vedlegg 1 Fremdriftsplan for etablering av testsett



Vedlegg 2 Tillatelser og annen korrespondanse

- Brev ”Melding av kvalitetsssikringsprosjekt: Etablering av testsett for radiologer som tyder mammografibilder” og protokoll for arbeidet ble sendt 2. april 2012, etterfulgt av meldeskjema 2. mai 2012.
- Personvernombudet gir sin tilråding til innsamling og behandling av personopplysninger for intern kvalitetssikring ”*Etablering av testsett for radiologer som tyder mammografibilder*” 21. mai 2012.
- Brev til medisinsk faglig ansvarlig (radiolog) og ledende radiograf ble sendt 23. april 2012 til fire Brystdiagnostiske sentra med forespørsel om uthenting av bilder.
- Notat av 31.mai 2012: Identifikasjon og uttrekk av screeningmammogrammer – en spesifikasjon for uttrekk i mammaseksjonen, Kreftregisteret.
- Veiledning av 20. juni 2012: ”Mammografi testsett - uthenting av mammogrammer ved Haukeland”.
- Abstract ble sendt 15.april 2012 til International Cancer Screening Network (ICSN) symposium i Sydney, Australia okt 2012. Antatt for poster presentasjon 25. mai 2012. Poster ble presentert på ICSN av veileder S. Hofvind 24. oktober 2012.

Vedlegg 3 Screeningmammogrammer i testsettene

Tabellene under viser *løpenummer* og resultat av utrekk og gjennomgang av screeningmammogrammene i testsettene. Tallene i kursiv er løpenummer, de øvrige er antall screeningundersøkelser.

Testsett GE (Haukeland) har *løpenummer* fra 100 til 219.

Vedlegg 3a Testsett GE (Haukeland)

Testsett 1	**GE**	50-54	55-59	60-64	65-69
SN ønsket	80	22	21	16	21
SN uttrekk	88	24	23	18	23
SN benyttes	74	*100,101, 104,114, 119,124, 127,128, 135,137, 144,150, 158,165, 166,172, 177,181, 182,219*	*117,118,126, 129,130,138, 140,156,163, 164,174,180, 184,187,193, 201,203,204, 211*	*113,122,123, 125,133,153, 154,161,169, 175,179,196, 197,208,210, 214*	*105,107, 108,120, 121,131, 132,148, 151,152, 170,188, 189,194, 195,199, 205,206, 212*
SN eksluderes	13	*167,171, 176,185*	*110, 186,192,*	*106,168*	*109,141, 146,190*
SN benyttes	**61**	**16**	**16**	**14**	**15**
SP ønsket	20	5	5	5	5
SP uttrekk	32	8	8	8	8
SP benyttes	27	*139,145, 159,198, 207,218*	*111,134,136, 147,191,200, 209*	*103,157,160, 173,213*	*102,112, 115,116, 143,149, 162,215*
SP eksluderes	5	*178,217*	*183*	*142,155*	
SP eksluderes v/regranskning	1			*202*	
SP benyttes	**26**	**6**	**7**	**5**	**8**
Undersøkelser	*100*	*27*	*26*	*21*	*26*
I testsett	*100*	*26*	*26*	*21*	*27*

Testsett Hologic (Vestfold) har *løpenummer* fra 300 til 419.

Vedlegg 3b Testsett 2 Hologic (Vestfold)

Testsett 2	**Hologic**	50-54	55-59	60-64	65-69
SN ønsket	85	24	22	17	22
SN uttrekk	93	26	24	19	24
SN kan benyttes	88	*300,301, 312,317, 323.326, 331,334, 344,355, 358,362, 363,364, 367,381, 384,389, 390,391, 394,399, 415,416*	*306,308,314, 315,321,339, 345,346,352, 353,361,372, 376,377,385, 386,387,396, 398,403,406, 410,412*	*310,327,328, 332,340,350, 360,366,368, 370,371,395, 397,400,405, 407,411*	*309,316, 318,320, 329,330, 335,343, 347,348, 351,365, 369,373, 374,378, 379,382, 388,392, 393,402, 404,408*
SN eksluderes	5	*333,349*	*319*	*324,359*	
SN benyttes	**85**	**24**	**de 22 første**	**17**	**de 22 første**
SP ønsket	15	6	3	3	3
SP uttrekk	27	9	6	6	6
SP benyttes	24	*302,303, 307,311, 313, 356, 380,413*	*336,342,354, 357,409,**414***	***304**,322,325, 337,338,419*	*305,341, **375**,383, 418*
SP eksluderes	2	*401*			*417*
SP eksluderes v/regranskning	8	*313 2 lesj, 413 funn cc*	*342 mfokal, 354 side?,409, 414 finner ikke*	*325 – 2 foci*	*383 bilat*
SP benyttes	**15**	**302,303,307 ,311,356, 380**	**336,357**	**304,322,337, 338**	**305,341, 375,**
Undersøkelser	*100*	*30*	*25*	*20*	*25*
I testsett	*100*	*30*	*25*	*20*	*25*

Testsett Philips Sectra (Trøndelag St Olav) har *løpenummer* fra 500 til 619.

Vedlegg 3c Testsett 3 Philips Sectra (Trøndelag St Olav)

Testsett 3	**Philips Sectra**	50-54	55-59	60-64	65-69
SN ønsket	85	24	22	17	22
SN uttrekk	93	26	24	19	24
SN benyttes	92	*500,501, 506507, 509,511, 518,522, 524,525, 532,533, 535,539,544 557,558,559 573,575, 585,586,593 ,596,600, 605*	*505,514,517, 526,529,536, 537,540,541, 553, 563,564, 565,567,571, 579,582,583, 592,594,611, 613,615,618*	*502,508,530, 546,561,562, 566,584,589, 590,591,595, 597,598,599, 601,608,614*	*503,512, 519,523, 531,538, 545,547, 548,549, 550,551, 556,560, 568,569, 570,572, 574,604, 607,610, 616,617*
SN eksluderes	1			516	
SN benyttes	**85**	**de 24 første**	**de 22 første**	**de 17 første**	**de 22 første**
SP ønsket	15	6	3	3	3
SP uttrekk	27	9	6	6	6
SP benyttes		*504,510, 521,528,534 ,554,**577**, 580*	*520,**542**,576, 609,612,619*	*513,515,552, 578,587, 588*	*527,555 ,**581**,602, 603,606*
SP eksluderes	1	*543*			
SP eksluderes v/regranskning	3	*504, 521 2 lesj*		*552 bilat*	
SP benyttes	**15**	***510,528,534 ,554,577,58 0***	*520,**542**,576*	*513,515,578*	*527,555, **581***
Undersøkelser	*100*	*30*	*25*	*20*	*25*
I testsett	*100*	*30*	*25*	*20*	*25*

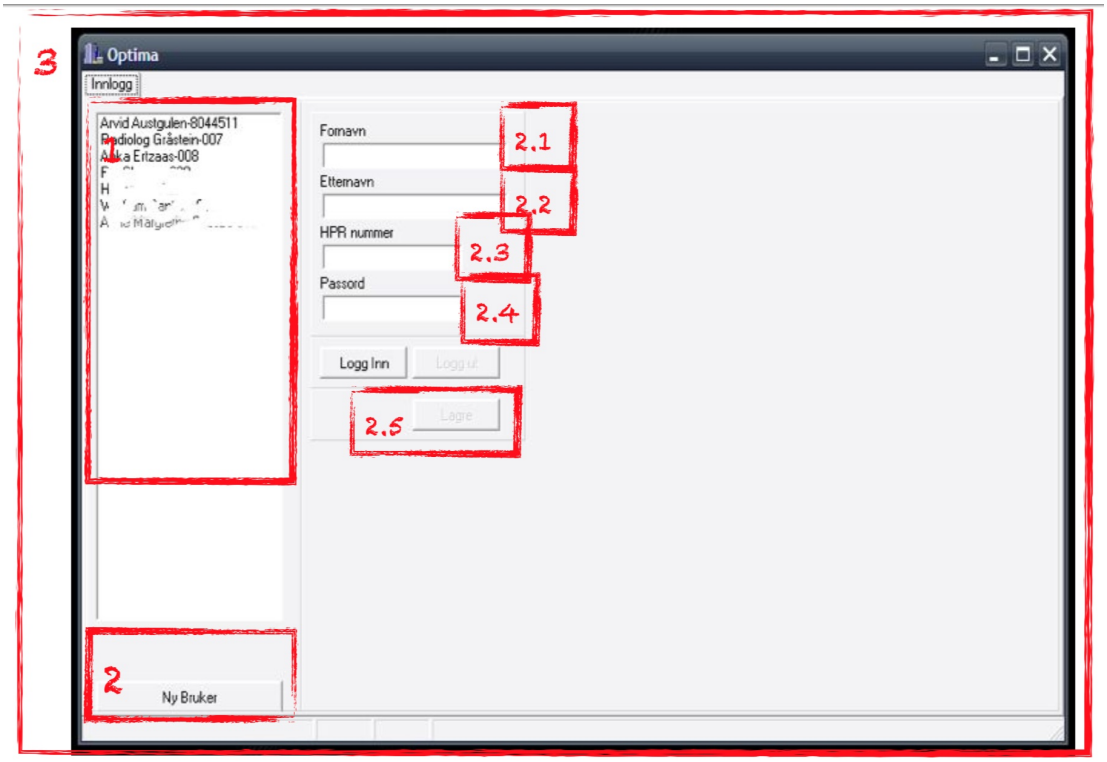
Testsett Siemens (Vestre Viken, Drammen) har *løpenummer* fra 700 til 819.

Vedlegg 3d Testsett 4 Siemens (Vestre Viken, Drammen)

Testsett 4	**Siemens**	50-54	55-59	60-64	65-69
SN ønsket	75	22	21	16	16
SN uttrekk	83	24	23	18	18
SN på hard-disk	76	*700,709,711, 723,730,732, 733,736,741, 745,758,759, 768,781,784, 788,789,797, 802,806,808, 809, 818*	*705,707,712, 717,718,719, 729,734,737, 752,757,766, 768,769,772, 783,793,800, 805,807,813*	*703,710,722, 724,728,744, 748,751,762, 764,773,775, 787,792,794, 799, 801*	*704,708, 716,725, 731,753, 754,761, 763,765, 767,774, 779,815, 819*
SN eksluderes etter uttrekk	8	*778*	*743,760*	*738,756*	*706,747, 796*
SN benyttes		**de 22 første**	**21**	**de 17 første**	**15**
SP ønsker	25	7	6	6	6
SP uttrekk	37	**10**	9	9	9
SP på hard disk	36	***720**,735,739, 740,746,755, **777**,**780**,**798**, 811*	*713,**714**,771, 782,786,791, **814,816***	*715,726,727, 742,**749**,764, **785**,803,817*	*701,702, **721**,776, 790,795, 804,810, 812*
SP eksluderes etter uttrekk	1		750		
SP eksluderes v/regranskning		Ikke regransket			
SP benyttes					
Undersøkelser	*100*	*29*	*27*	*22*	*22*
I testsett					

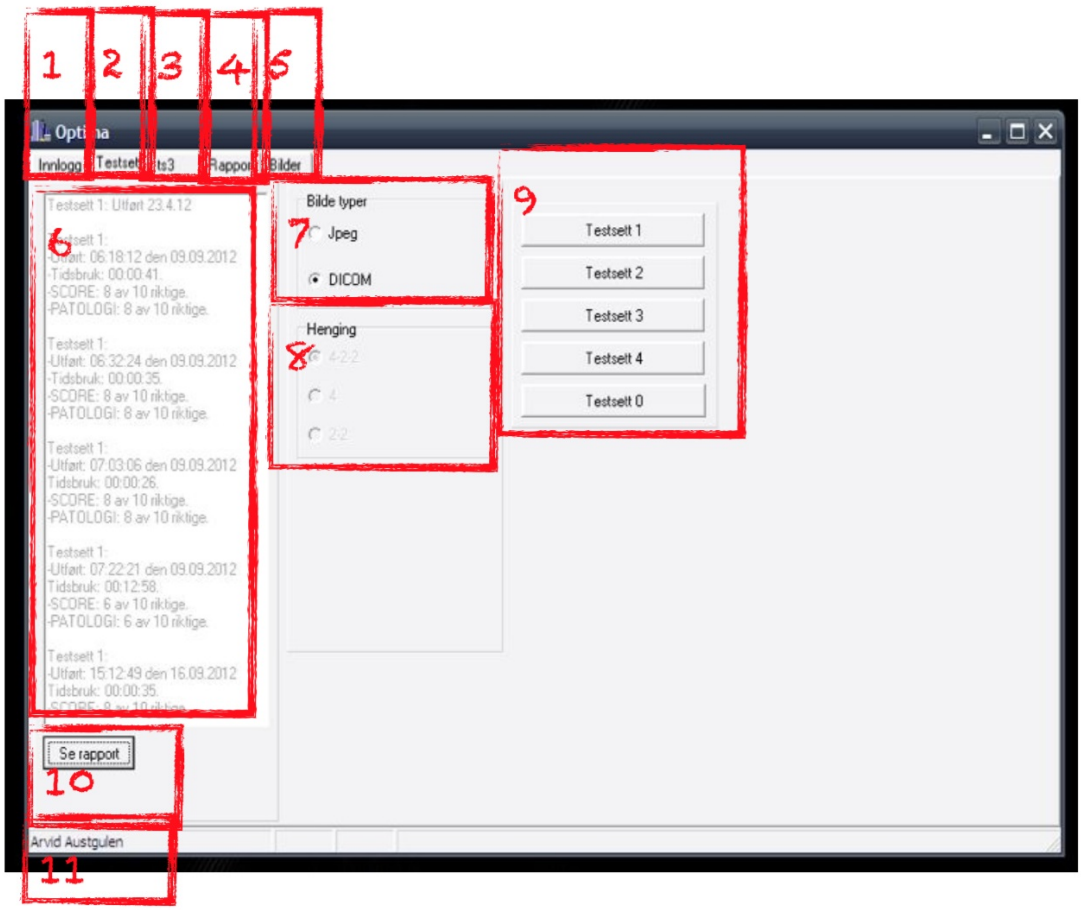
Vedlegg 4 Skjermbilder

Vedlegg 4a Innlogging 1



1. Liste over allerede registrerte brukere i systemet for aktuell terminal. Man velger bruker ved å klikke på aktuelt navn.
2. Nye brukere registrerer seg her.
   - 2.1 Man skriver inn fornavn.
   - 2.2 Man skriver inn etternavn.
   - 2.3 Man skriver inn sitt HPR-nummer.
   - 2.4 Man skriver inn selvvalgt passord.
   - 2.5 Man klikker ”Lagre”.
3. Hovedvinduet til applikasjonen.

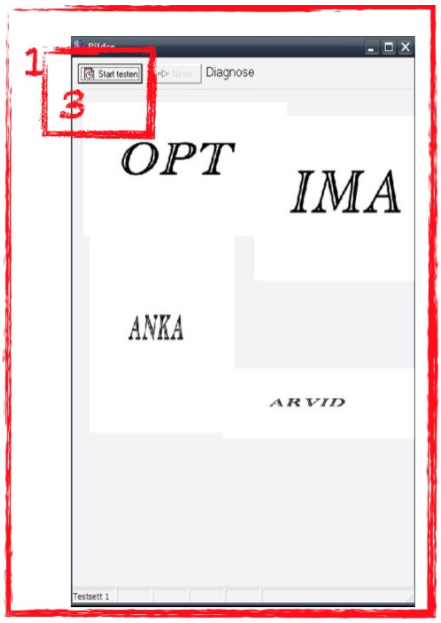
Vedlegg 4b oversikt over egne tester

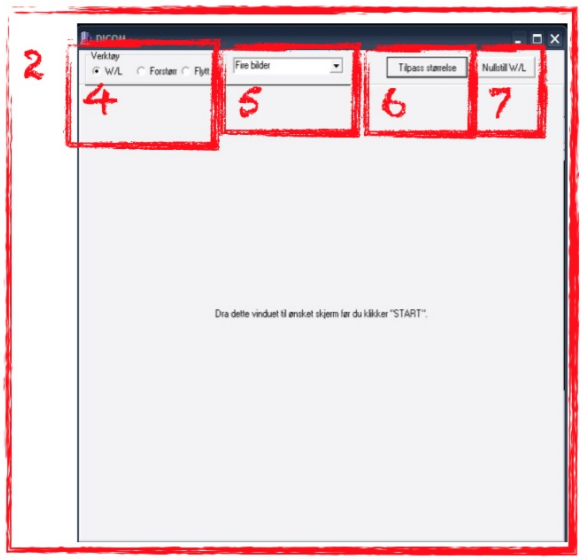


Nye faner dukker opp etter innlogging.

1. Innloggings- og utloggings-fane.
2. Testsett-fane.
3. Administrator-fane som kun er synlig for dem som er registrerte som administratorer av applikasjonen.
4. Rapport-fane hvor man kan åpne lagrede rapporter fra tidligere tester.
5. Bilde-fane hvor man etter en fullført test kan åpne aktuelle bilder.
6. Oversikt over tidligere utførte tester på innlogget bruker.
7. Valg man kan gjøre mellom å JPEG-bilder eller DICOM-bilder. Testsettene er i oppgaven bare lagt inn i DICOM-formatet.
8. Valg av hengeprotokoller.
9. Valg av hvilket testsett man ønsker å ta.
10. Knapp for å gå til rapport-fanen.
11. Innlogget bruker.

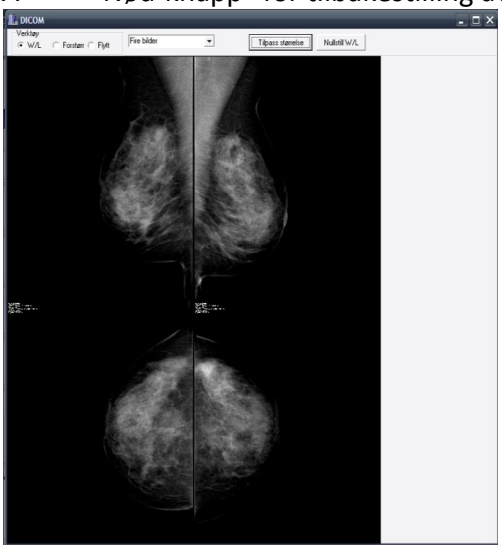
Vedlegg 4c Start av test



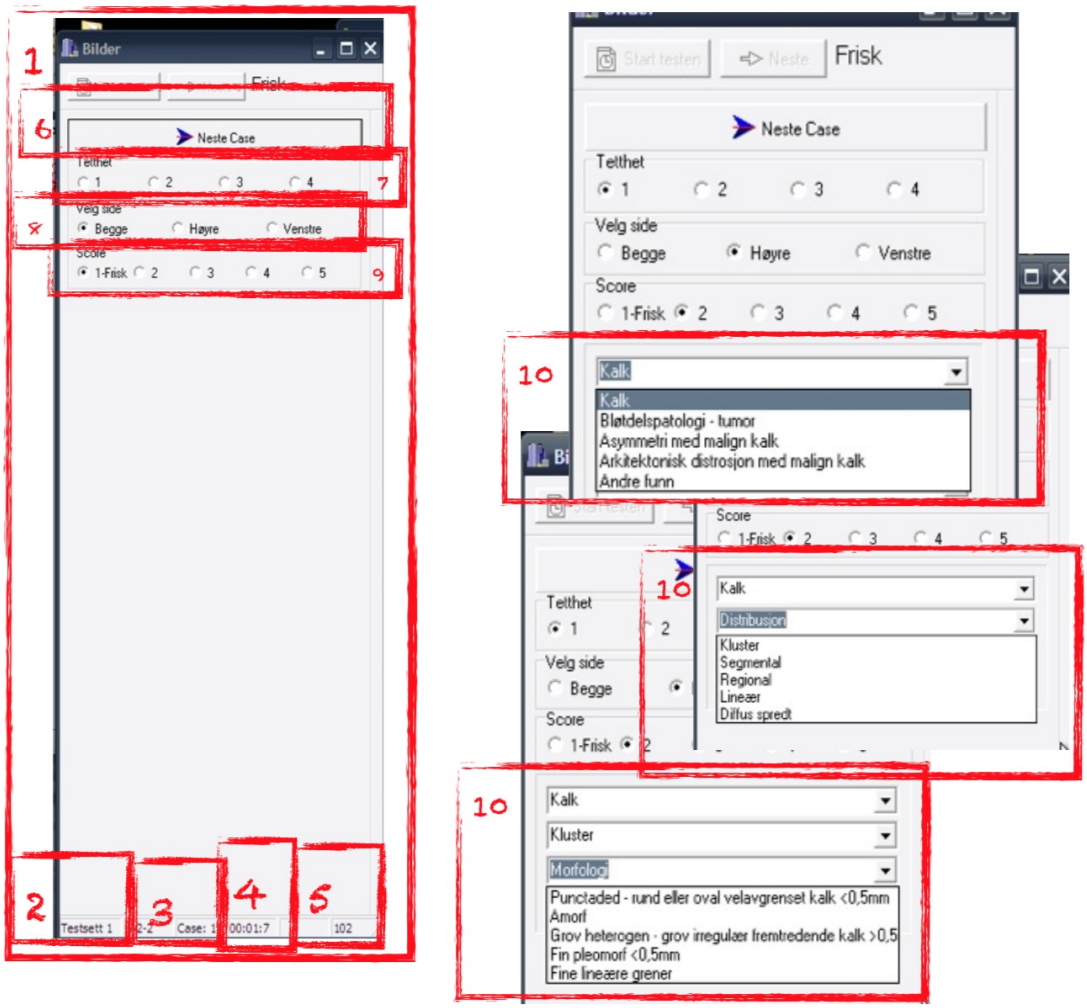


Når man starter en test vil to nye vinduer dukke opp.

1. Diagnose-vinduet hvor man setter om case er “frisk” eller angir diagnose.
2. Bilde-vindu, hvor man ser på bildene når testen startes.
   Dette vinduet føres over til aktuell skjerm hvis man har flere enn én.
3. Knapp for strat av testen.
4. Her kan man velge mellom de tre verktøyene man har.
   ”W/L” hvor man kan styre window og level med musen.
   ”Forstørr” hvor man får et forstørrelsesglass i stedet for musepeker.
   ”Flytt” hvor man kan flytte på bildet ved stor forstørrelse.
5. Rullegardin-meny hvor man kan velge å forstørre ett bilde til hele skjermen.
6. ”Nød-knapp” for tilpassing av billedstørrelser.
7. ”Nød-knapp” for tilbakestilling av window og level.

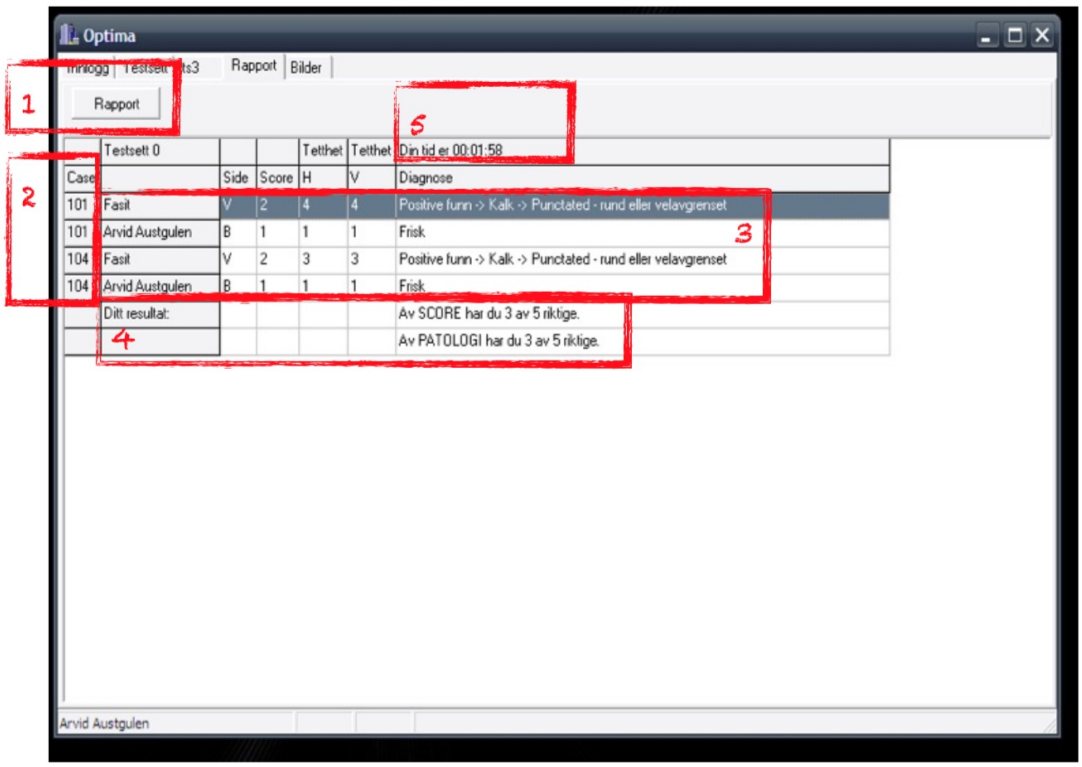


Vedlegg 4d Diagnose-vindu



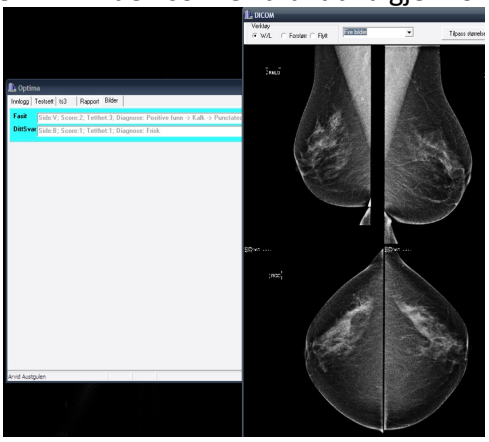
1. Diagnosevindu.
2. Viser hvilket testsett som kjøres.
3. Viser hvilken case man ser på fra 1 til 100.
4. Viser tiden man har brukt så langt i testsettet
5. Viser faktisk case-nummer.
6. Knapp for neste case.
7. Tester angir tetthet her.
8. Tester angir side for patologi her, eller ”begge” ved ”frisk”.
9. Tester angir score her.
10. Ved score >1 eller side ulikt ”begge” vil nye valg dukke opp for testeren.

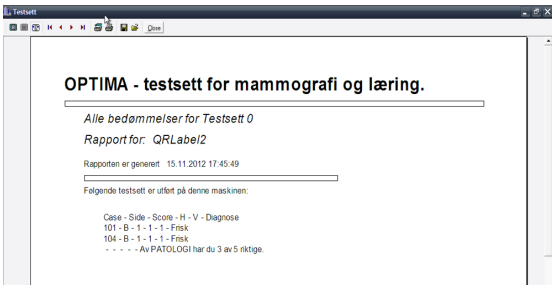
Vedlegg 4e Resultat



Etter at testen er ferdig vises rapporten over alle svar som fraviker fasiten samt alle patologiske svar.

1. Knapp for rapport som man kan skrive ut eller lagre (se bildet under til høyre).
2. Nummeret til aktuell case.
3. Viser hvem svaret i raden tillhører; fasit eller testeren.
   Ved å dobbel-klikke på en av disse linjene vises fasit og resultat samt tilhørende bilder på valgt case (se bilde under til venstre)
4. Resultatet oppsummert vies til slutt.
5. Tiden som er brukt til å gjennomføre testsettet.





Vedlegg 5 Parameterutvalg i registreringsløsning

Parameterutvalget i programvaren ved andre involvering

- Frisk
- Positivt funn
    - Kalk
        - Punctated – rund eller oval velavgrenset kalk <0,5 m.m.
            - Grupert eller kluster
            - Segmental
            - Regional
            - Lineær
            - Diffus
        - Amorphous –
            - Grupert eller kluster
            - Segmental
            - Regional
            - Lineær
            - Diffus
        - Grov heterogen > grov irregulær fremtredende kalk på >0,5 m.m.
            - Grupert eller kluster
            - Segmental
            - Regional
            - Lineær
            - Diffus
        - Fin Pleomorph < 0,5 m.m.
            - Grupert eller kluster
            - Segmental
            - Regional
            - Lineær
            - diffus
        - Fine lineære greiner
            - Grupert eller kluster
            - Segmental
            - Regional
            - Lineær
            - Diffus
    - Svulst/Tumor/masse
        - Form
            - Oval
            - Rund
            - Irregulær
            - Lobulær
        - Avgrensning

                - Velavgrenset
                - Indestingt
                - Spikulert
            - Tetthet
                - Høy
                - Isodent
                - Lav
    - Asymmetri
    - Distorsjon
    - Masse med kalk
        - Alle over
    - Asymmetri med kalk
        - Kalk som over
    - Distrosjon med kalk
        - Kalk som over

Parameterutvalg ved tredje involvering (reganskning)

Positivt funn

- **Kalk**
    - Distribusjon
        - Kluster
        - Segmental
        - Regional
        - Lineær
        - Diffus spredt
    - Morfologi
        - Punctated – rund eller oval velavgrenset kalk < 0,5 m.m.
        - Amorf
        - Grov heterogen > grov irregulær fremtredende kalk på > 0,5 m.m.
        - Fin pleomorf < 0,5 m.m.
        - Fin lineære greiner
- **Bløtdelspatologi - tumor**
    - Form
        - Oval
        - Rund
        - Lobulær
        - Irregulær
    - Avgrensning
        - Velavgrenset
        - Skjult- utvisket
        - Uskarp
        - Spikulert

        - Tetthet
            - Høy
            - Isodent
            - Lav
    - **Asymmetri med malign kalk**
    - **Arkitektonisk distorsjon med malign kalk**
    - **Andre funn**

Vedlegg 6 Brukervennlighetsskala

Tabellene i *Vedlegg 6a-6e* viser resultater fra brukbarhetstesting ved regranskning og utprøving av testsett.

Vedlegg 6a Resultat SUS test ved regranskning

Spørsmål	**Sterkt uenig**				**Enig**	**Poeng**
	1	**2**	**3**	**4**	**5**	
Jeg tror jeg vil benytte testsettene					x	4
Jeg mener programmet er unødvendig komplekst	x					4
Programmet er lett å bruke					x	4
Jeg trenger veiledning for å kunne bruke programmet	x					4
De ulike funksjonene er godt integrert			x			2
Det er for mye inkonsistens i programvaren		x				3
Jeg tror de fleste vil lære å bruke programmet svart raskt					x	4
Jeg synes det er tungvint å bruke systemet	x					4
Jeg følte meg trygg på å bruke programmet					x	4
Jeg trengte å lære mye før jeg kunne komme i gang med programmet	x					4
Poeng						37
SUM SUS score (poeng*2.5)						**92.5**

Vedlegg 6b Resultat SUS test, Radiolog 1

Spørsmål	**Sterkt uenig**				**Enig**	**Poeng**
	1	**2**	**3**	**4**	**5**	
Jeg tror jeg vil benytte testsettene		x				1
Jeg mener programmet er unødvendig komplekst	x					4
Programmet er lett å bruke				x		3
Jeg trenger veiledning for å kunne bruke programmet				x		1
De ulike funksjonene er godt integrert				x		3
Det er for mye inkonsistens i programvaren		x				3
Jeg tror de fleste vil lære å bruke programmet svart raskt					x	4
Jeg synes det er tungvint å bruke systemet			x			2
Jeg følte meg trygg på å bruke programmet					x	4
Jeg trengte å lære mye før jeg kunne komme i gang med programmet	x					4
Poeng Regransker 1						29
SUM SUS score (poeng*2.5)						**72.5**

Regransker en ga programvaren en brukbarhetsscore på 72,5 poeng

Vedlegg 6c Resultat SUS test, Radiolog 2

Spørsmål	**Sterkt uenig**				**Enig**	**Poeng**
	1	**2**	**3**	**4**	**5**	
Jeg tror jeg vil benytte testsettene				x		3
Jeg mener programmet er unødvendig komplekst	x					4
Programmet er lett å bruke					x	4
Jeg trenger veiledning for å kunne bruke programmet	x					4
De ulike funksjonene er godt integrert			x			2
Det er for mye inkonsistens i programvaren		x				3
Jeg tror de fleste vil lære å bruke programmet svart raskt					x	4
Jeg synes det er tungvint å bruke systemet	x					4
Jeg følte meg trygg på å bruke programmet					x	4
Jeg trengte å lære mye før jeg kunne komme i gang med programmet	x					4
Poeng Regransker 2						37
SUM SUS score (poeng*2.5)						**92.5**

Regransker to ga programvaren en brukbarhetsscore på 92,5 poeng

Vedlegg 6d Resultat SUS test, Radiolog 3

Spørsmål	**Sterkt uenig**				**Enig**	**Poeng**
	1	**2**	**3**	**4**	**5**	
Jeg tror jeg vil benytte testsettene					x	4
Jeg mener programmet er unødvendig komplekst	x					4
Programmet er lett å bruke					x	4
Jeg trenger veiledning for å kunne bruke programmet			x			2
De ulike funksjonene er godt integrert					x	4
Det er for mye inkonsistens i programvaren	x					4
Jeg tror de fleste vil lære å bruke programmet svart raskt					x	4
Jeg synes det er tungvint å bruke systemet	x					4
Jeg følte meg trygg på å bruke programmet					x	4
Jeg trengte å lære mye før jeg kunne komme i gang med programmet	x					4
Poeng Regransker 3						38
SUM SUS score (poeng*2.5)						**95**

Regransker tre ga programvaren en brukbarhetsscore på 95 poeng

Vedlegg 6e Resultat SUS test, Radiolog 4

Spørsmål	**Sterkt uenig**				**Enig**	**Poeng**
	1	**2**	**3**	**4**	**5**	
Jeg tror jeg vil benytte testsettene				x		3
Jeg mener programmet er unødvendig komplekst			x			2
Programmet er lett å bruke			x			2
Jeg trenger veiledning for å kunne bruke programmet					x	0
De ulike funksjonene er godt integrert					x	4
Det er for mye inkonsistens i programvaren	x					4
Jeg tror de fleste vil lære å bruke programmet svart raskt					x	4
Jeg synes det tungvint å bruke systemet			x			2
Jeg følte meg trygg på å bruke programmet		x				1
Jeg trengte å lære mye før jeg kunne komme i gang med programmet			x			2
Poeng Regransker 4						24
SUM SUS score (poeng*2.5)						**60**

Regransker fire ga programvaren en brukbarhetsscore på 60poeng

Vedlegg 7 Poster

Poster som ble presentert på International Cancer Screening Network (ICSN) symposium i Sydney, Australia 24. Oktober 2012

Establishing test sets for measuring reader performance in the Norwegian Breast Cancer Screening Program

Ertzaas AK*, Austgulen A** and Hofvind S*.
The Cancer Registry of Norway, Oslo * and Curato X-ray, Bergen **

BACKGROUND

The effectiveness of mammographic screening depends on the radiologist's ability to interpret the screening mammograms. A key issue is to maintain and improve the interpretation skills.

Test sets of screening mammograms might be one way to increase the sensitivity of the reader performance as a supplement to the recommended quality assurance. Hence, we have created four test sets of screening mammograms for this purpose.

METHODS

Each set of 100 screening mammograms contains mammograms from one supplier; General Electric, Hologic, Phillips Sectra or Siemens. Screening examinations including true negative (85-75%) and true positive cases (25-15%) are randomly selected from the national screening database. All mammograms are obtained from a hospital PACS and thereafter anonymized.

Radiologists contributed in the planning, testing and finilizing the system. We have used a "System Usability Scale" to obtain the radiologists opinion of the software usability.

DISCUSSION

To maintain and improve the radiologists skills is a key issue in mammographic screening. One way to do this is to participate in a regular educational self-assessment scheme. The test set will work as an educational self-assessment and training scheme for breast screening professionals. The test sets can be used as a tool to increase the sensitivity of the reader performance and in clinical audits.

REGISTRATION SYSTEM

A registration system was developed to collect information about the radiologists interpretation of the screening mammograms and classification of the positive cases.

Options related to zoom, window and level during interpretation.

Information collected:

General information:
- Date of interpretation
- Time used for the interpretation
- Are the test set read earlier
- Identification of test set number

Registered by radiologist:
- UserID
- Mammographic density (BI-RADS)
- Selected for further assesment: No/yes

Mammographic findings (modified BI-RADS)
- Calcification (distribution – morphology)
- Masses (shape – margin – density)
- Asymmetri
- Distortion

REPORT MODUL

Immediate feedback of the interpretation and classification.

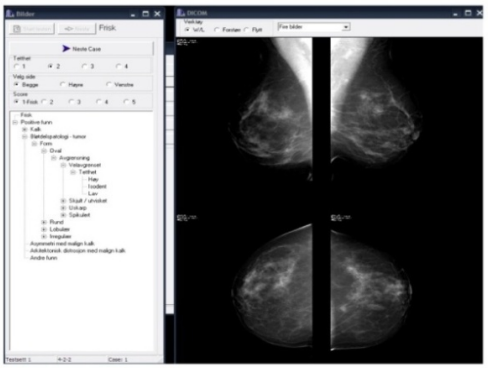
Example:

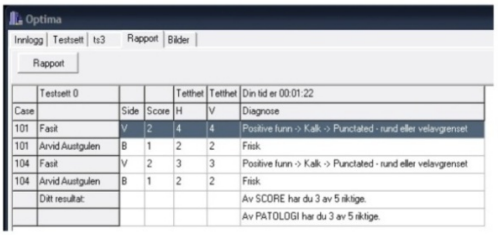
Report for Dr Read Super

Test set 1 read: Date, time and time used
# of correct interpretations
# of correct classifications

Opportunity to return to mammograms with incorrect interpretation and/or classification.

SCREENSHOTS





Corresponding author: anne.kathrin.ertzaas@kreftregisteret.no