Vitenskapelige kriterier for innføring og evaluering av screening

Sammendrag

Bakgrunn. Effekten av screening på dødelighet dokumenteres best i randomiserte studier, men mange screeningprogram er innført uten slike. Screening medfører alltid en insidensøkning, og den kan være meget stor. Hvis det kommer en bedre behandlingsmetode eller en ny og bedre diagnostisk metode, så vil effekten av screening forandres. Derfor bør man alltid studere insidens og dødelighet i oppfølgingsstudier.

Materiale og metode. Reduksjon i dødelighet ved screening illustreres her med svenske og norske tall for bryst- og livmorhalskreft. Effekten av variabel veksthastighet på screening (heterogenitet) diskuteres. Litteratursøk er gjort med søkeordene «screening», «mammography screening» og «cervix screening» i Medline.

Resultater. 10–15 år med mammografiscreening i Sverige er assosiert med 10 % reduksjon i dødelighet av brystkreft for aldersgruppen 50–69 år. Mammografiscreening i Sverige er assosiert med 50 % vekst i insidens. Pap-screening (Papanicolaou-teknikk, utstryk) for livmorhalskreft har redusert dødeligheten med 30 % de siste 20 år i både Norge og Sverige.

Fortolkning. Kreftscreening basert på tidligere diagnostikk har noen ganger liten effekt på dødelighet. Dette kan skyldes at man ved screening primært oppdager saktevoksende svulster, og mange av disse ville ellers aldri blitt oppdaget. Reduksjon i dødelighet ved screening må avveies mot negative sider. De viktigste er overdiagnostisering, overbehandling, falskt positive og falskt negative prøver.

I Tidsskriftet nr. 3–5/2003 publiseres en serie artikler som viser veien fra forskning til hverdagsmedisin. Serien er initiert av Jahn M. Nesland.

Engelsk sammendrag finnes i artikkelen på *www.tidsskriftet.no*

> Se også side 301

Per-Henrik Zahl
per-henrik.zahl@fhi.no
Avdeling for helsestatistikk
Nasjonalt folkehelseinstitutt
Postboks 4404
0403 Nydalen

Screening betyr masseundersøkelse av symptomfrie mennesker og sortering av individer i to grupper med henholdsvis lav og høy risiko for å bli eller være syk (silingsundersøkelse) (1). Individer med høy risiko må ta en ny prøve før man kan si sikkert om de har en sykdom eller vil få den. Undersøkelse av celleforandringer (Pap-test) som kan føre til livmorhalskreft, er et eksempel på masseundersøkelse som faller inn under definisjonen av screening.

Formålet med screening er å stille diagnosen på et tidlig stadium. Hvis man fjerner alle kreftceller før de sprer seg, eller forstadier før de utvikler seg til kreft, er prognosen bedre. Man kan bruke røntgenundersøkelse, en enkel laboratorietest eller ren visuell inspeksjon. Mange av testene brukes således i vanlig klinisk praksis for å verifisere mistanke om sykdom, men noen er utviklet spesielt for screening, f.eks. Pap-test. Ved Pap-screening diagnostiseres og behandles mange individer som ikke ville utviklet livmorhalskreft. Samtidig er det mange individer med celleforandringer som ikke blir oppdaget (2).

Det er generell enighet om at randomiserte forsøk med høy deltakelse kan gi et valid estimat på nytten av screening (1, 3). Påvist effekt i randomiserte studier bør være en nødvendig betingelse for at man skal starte med organisert screening i en populasjon (3). Observert effekt av screening i det virkelige liv kan være vesentlig forskjellig fra effekten i randomiserte studier. Dette kan enten forklares med at det er kommet bedre diagnostiske metoder eller bedre behandling. Det kan også forklares med at det var vesentlige feil i de randomiserte studiene. Det er derfor vanlig å kreve at effekten av screening skal dokumenteres i oppfølgingsstudier (3).

Screening har både positive og negative sider, og samlet effekt er ikke alltid positiv for pasienten. De viktigste negative sidene ved screening er overdiagnostisering, falskt negative prøver og falskt positive prøver. Mammografiscreening har ført til at antall kvinner som behandles for brystkreft eller forstadier til brystkreft (begge får samme behandling) har økt med 70 % på ti år i aldersgruppen 50–69 år i Sverige, men knapt i andre aldersgrupper (4). Det kan også tenkes at denne økningen delvis skyldes en reell økning i forekomst. Falskt negative mammografier er i dag vanligste årsak til erstatningssøksmål mot helsevesenet i USA (5).

! Fakta

Forslag til kriterier

Screening bør dokumenteres og evalueres på samme måte som legemidler før det innføres som et allment helsetilbud.

- Positive sider (reduksjon i dødelighet) bør være klart større enn de negative sider (overdiagnostisering, overbehandling, falskt negative prøver, falskt positive prøver).
- Reduksjon i dødelighet bør være påvist i flere randomiserte studier og i flere land.
- Data om hvilke individer som er ekskludert, bør være offentlig tilgjengelig. Validitet i randomiserte studier står og faller på at det ikke har vært selektiv eksklusjon av individer.
- Det har vist seg at negative sider ved screening blir systematisk underrapportert (11, 21). Det er derfor nødvendig at nytten av screening blir evaluert av et uavhengig organ.

Definisjoner

Når man screener for sjeldne sykdommer med en metode som er dårlig i den forstand at man ikke oppdager alle som er syke og samtidig plukker ut mange flere for videre testing enn de som virkelige er syke, skapes det en rekke problemer.

De som tester positivt i én screening, men ikke har sykdommen, sier man har en *falskt positiv test*. Ved mammografiscreening anbefales det å ta 20 bilder i løpet av perioden 40–70 år (6). Hvis sannsynligheten for en falskt positiv prøve er 5 %, vil muligheten for at et individ skal oppleve en eller flere falskt positive prøver etter 20 tester være $1 - (0{,}95)^{20} = 0{,}65$. Dvs. hele 65 % vil opp-

leve en eller flere falskt positive tester i løpet av en periode med 20 tester. De som tester negativt, men som er syke, har en *falskt negativ test*.

Sensitiviteten av en test er sannsynligheten for at en person med sykdommen skal teste positivt. *Spesifisiteten* er sannsynligheten for at en person uten sykdommen skal teste negativt. Sensitiviteten til Pap-tester er anslått til 10–90 % (2), og sensitiviteten til mammografi anslås til 54–94 % (varierer med alder) (7).

Overlevelsestid er tid fra diagnose til død eller et annet sluttpunkt. *Fremskyndingstid* (lead-time) er den tiden som diagnosen fremskyndes ved screening. Dette er et sentralt mål ved evaluering av screening. Fremskyndingstid er i seg selv første indikator på effekten av et screeningprogram. Insidensraten stiger umiddelbart etter start av screening fordi man oppdager mange små svulster som man ville oppdaget på et senere tidspunkt.

Heterogenitet og seleksjon av saktevoksende/lavmaligne svulster

Mange sykdommer, for eksempel kreft, består av en blanding av alvorlige og mindre alvorlige varianter, der noen utvikler seg fort og har høy dødelighet, mens andre varianter er harmløse og har lav dødelighet. Denne forskjellen i dødelighet kalles for heterogenitet. Ofte har man lite kunnskap om hvilke svulster som er mest dødelige ved diagnosetidspunktet, og gjerne mindre kunnskap jo tidligere man stiller diagnosen. Selv om man ved diagnosetidspunktet ikke kjenner hvilke individer som har høyest dødelighet, er det viktig å forstå hvilken effekt heterogenitet har på effekten av screening.

Hvis mammografiscreening fremskynder diagnosen med én doblingstid (gjennomsnittlig doblingstid for brystkreftsvulster i en studie var 82 dager (8)) og kvinner testes hvert annet år, vil bare noen få av disse bli oppdaget tidligere ved screening. Svulster med doblingstid på over to år vil alltid oppdages tidligere. Hvis man undersøker alle individer med samme mellomrom, har personer med saktevoksende svulster relativt større sannsynlighet for å bli oppdaget ved screening enn personer med rasktvoksende svulster. Dette fenomenet kalles her for *lavmalign seleksjon* (length-time bias) (1). Individer som oppdages mellom to screeningtester, har ofte dårligere prognose enn de som oppdages ved screening (9). Deres sykdommer omtales ofte som intervallkreft og er typisk høymaligne.

Tidligere diagnose vil også medføre at man oppdager mange små og saktevoksende svulster som ellers aldri ville blitt diagnostisert (overdiagnostisering) (1). Lavmalign seleksjon og overdiagnostisering er derfor to fenomener som henger sammen med heterogenitet. Effekten av heterogenitet på screening er lite studert, men fenomenet er sentralt for å kunne forstå dynamiske forandringer i insidens- og dødelighetsrater som følge av screening.

Randomiserte studier

Det finnes i prinsippet to forskjellige typer vitenskapelige studier for å vurdere effekten av screening (1, 3): Eksperimentelle studier (randomiserte kliniske studier, intervensjonsstudier) og ikke-eksperimentelle (person-kontroll-studier, kohortstudier, økologiske studier mfl.). Hvis man sammenlikner overlevelsestid for to grupper hvorav den ene får tilbud om screening, vil forskjellen være systematisk skjev, fordi screening fremskynder diagnosetiden i den gruppen som får dette. I tillegg oppdager man ved screening relativt flere lavmaligne svulster. Randomiserte studier er en bedre vitenskapelig metode for å dokumentere nytten av screening enn ikke-eksperimentelle studier, fordi man studerer antall dødsfall fra studiestart og ikke overlevelsestid fra diagnosetidspunkt (1).

Ved studier av sjeldne sykdommer må man undersøke store populasjoner for å rekruttere mange nok syke individer til å starte et forsøk. Personer må følges over lang tid for å observere mange nok dødsfall til at man kan beregne en signifikant reduksjon i dødelighet med stor nok styrke. I den største svenske mammografistudien (WE-studien) deltok 135 000 kvinner, og de ble fulgt i sju år (10). I løpet av denne perioden ble det diagnostisert 1 663 brystkrefttilfeller, og det var 173 dødsfall pga. denne sykdommen. Kvinnene ble randomisert i én gruppe som fikk tilbud om screening med mammografi, og en kontrollgruppe som ikke fikk noe tilbud. I WE-studien ble kvinner randomisert gruppevis etter bosted, mens i andre studier er hvert enkelt individ blitt randomisert. Ved studiestart ble alle som var diagnostisert med kreft, ekskludert i de svenske studiene.

Det er i praksis umulig å gjøre screeningforsøk som er dobbeltblindt og innebærer 100 % oppfølging. For det første er det umulig å blinde deltakeren for en medisinsk prosedyre. Det har også vist seg vanskelig å blinde de personene som evaluerer utfallet av forsøket. For det andre er det alltid noen individer som flytter eller som man ikke klarer å følge opp og som må ekskluderes fra analysene. Dette medfører muligheter for flere systematiske feilkilder: Forskjellig eksklusjon av individer ved studiestart, ulik oppfølging og forskjellig klassifisering av dødsårsaker i studiegruppen og kontrollgruppen. I legemiddelindustrien gjøres forsøkene dobbeltblindt for å redusere muligheten for slike feil.

Det er vanlig å studere datavaliditet etter randomiseringen, slik man gjør i Cochrane-rapporter (11). Selektiv eksklusjon kan studeres ved å sammenlikne individene som er ekskludert i hver av gruppene med hverandre. I de svenske mammografistudiene ble alle som fikk diagnosen brystkreft etter studiestart, men før første mammografi i screeninggruppen, ekskludert (10, 11). Men data om hvem som ble ekskludert, er ikke offentlig tilgjengelig (11). Hvis man ekskluderer 10 % av kreftpasienter fra den ene gruppen, så vil man observere tilnærmet 10 % systematisk forskjell i dødelighet. Man kan også sammenlikne individer som ikke er ekskludert, f.eks. ved å studere aldersfordelingen blant dem som deltok (11).

Klassifisering av dødsårsaker er gjenstand for subjektiv vurdering. Hvis personene som klassifiserer dødsfallene har kunnskap om hvem som har fått tilbud om screening, er dette en kilde til systematiske feil. 10 % feilklassifisering av dødsfall gir 10 % systematisk lavere dødelighet.

Det er heller ikke opplagt at reduksjon i årsaksspesifikk dødelighet også gir reduksjon i totaldødelighet. For eksempel medfører mange former for kreftbehandling økt dødelighet av andre årsaker (12). Det er derfor ønskelig samtidig å studere reduksjon i totaldødelighet. Dette er i praksis umulig når man studerer sykdommer som brystkreft. Det ville kreve en meget stor studiepopulasjon (f.eks. 0,5 millioner) for å oppnå stor nok styrke til å påvise en reduksjon i dødelighet.

Overdiagnostisering

Ved screening blir noen individer diagnostisert før sykdommen har spredd seg, dermed blir effekten av behandling bedre. Samtidig diagnostiseres noen individer med svulster som vokser så sakte at de aldri ville blitt diagnostisert. Dette kalles her for overdiagnostisering. Et lite antall vil også bli diagnostisert før de rekker å dø av andre årsaker. Hvis tidligere diagnose ikke medfører bedre prognose, har det ingen hensikt å drive med screening.

Det er velkjent at screening ofte medfører en sterk forbigående økning i insidensrater i første runde (1, 3, 10). Etter andre runde skal forekomsten i gruppen som screenes, i teorien bare være marginalt høyere (1, 3). Når man slutter med screening, vil man se et fall i insidens (1). Dette betyr at ved mammografiscreening i Norge skal man se en liten økning i aldersgruppen 50–69 år og en liten reduksjon i aldersgruppen over 70 år hvis screening bare medfører tidligere diagnostikk av krefttilfeller som ellers ville blitt oppdaget. Totalt sett forventer man at antall diagnostiserte individer bare skal øke marginalt.

Det er imidlertid en erfaring at screening ofte medfører en vedvarende økning i insidens i den gruppen som screenes og ingen kompensatorisk reduksjon når individer ikke lenger screenes. Screening for brystkreft har medført en relativ økning i insidensrater på 50–60 % i Finland (13, 14), Sverige (4) og USA (15) i den aldersgruppen som screenes, og økningen har holdt seg i mer enn ti år etter at man startet med screening. Det har heller ikke kommet noe kom-

pensatorisk reduksjon i insidens i aldersgruppen over 70 år, som ikke screenes. Man ser ingen tilsvarende økning i land som ikke har startet med screening (16).

I figur 1 vises aldersspesifikke insidensrater for brystkreft i Sverige i periodene 1980–84 og 1995–99 (screening startet i 1987 og målgruppen er 50–69 år). Hvis økning i aldersgruppen 50–69 år skal forklares kun med tidligere diagnostikk, så skulle det ikke vært noen krefttilfeller i aldersgruppen 70–79 år.

Overdiagnostisering ved mammografiscreening er bare studert i en artikkel fra Australia (17) og er ellers bare kursorisk kommentert av Olsen & Gøtzsche (11). Overdiagnostisering er kanskje den viktigste negative siden ved screening for kreft. Mange forskere vegrer seg for å si at det man ser i figur 1 er overdiagnostisering (14) eller de har en annen forklaring (18). Vitenskapsjournalister i New York Times kaller dette for overdiagnostisering (19), og dette begrepet brukes også av statistikere (1, 3). Screening for nevroblastomer hos barn i Japan, Canada og Tyskland ble stanset fordi forekomsten av diagnostisert kreft steg med opptil 900 % (20, 21).

Evaluering

En generell regel ved evaluering av screening er at jo mindre reduksjon i dødelighet som observeres, desto viktigere blir det at man studerer og beregner de negative sidene. Toleransen for hva som er akseptable negative effekter av screening avhenger av i hvilken grad det resulterer i redusert dødelighet.

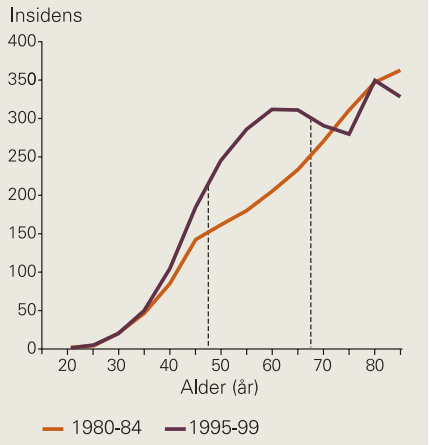
Mortalitetsstudier

Reduksjon i dødelighet er det endelige mål på nytten av screening. Det finnes en rekke grunner til at det er vanskelig å påvise eller estimere effekten av screening i befolkningsstudier. For det første forutsetter en generalisering av effekten i randomiserte studier at diagnostikk og behandling ikke har forandret seg. Stadiefordelingen av brystkreft har forandret seg, og dermed er fremskyndingstiden forandret de siste 20–40 år. Det har også kommet bedre behandling (22). Derfor kan man egentlig ikke si at effekten av screening i dag er den samme som i 20–40 år gamle randomiserte studier. For det andre kan det tenkes at risikofaktorene har forandret seg, men dette kan man i liten grad uttale seg om fordi disse er lite kjent. For det tredje – hvis organisert screening er begrenset til en liten del av befolkningen og det foregår uorganisert screening i resten av befolkningen, har man ingen naturlig kontrollgruppe for sammenlikning og effekten i kohortstudier blir underestimert.

I figur 2 presenteres dødelighetsrater av brystkreft i Sverige for perioden 1985–89 og 1995–99. Det skal ikke være noen effekt av screening før minst fire år etter start i 1987 (10). I praksis blir det meget vanskelig å påvise noen effekt av screening i slike data, og ingen har heller påvist en signifikant reduksjon i dødelighet som kan tilskrives mammografiscreening i noen kohortstudier (14).

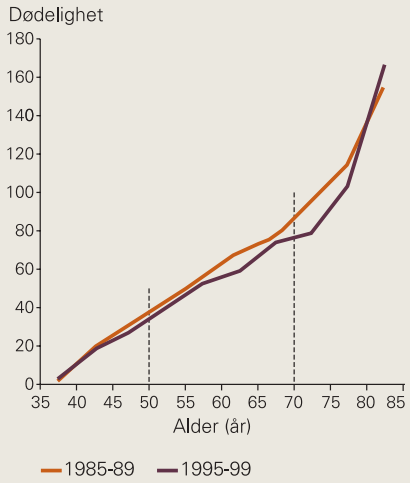
Man bør bruke konfidensintervaller når man skal generalisere effekten i randomiserte studier til en befolkning. En 10 % reduksjon i brystkreftdødelighet i Sverige i aldersgruppen over 50 år (fig 2) er ikke helt uforenlig med 10–30 % redusert dødelighet i randomiserte studier (11, 23). Det er mer problematisk at man ikke ser noen reduksjon i dødelighet i Finland (13, 24) og Nederland (24), der man startet med screening samtidig, men at brystkreftdødeligheten reduseres i Irland, Italia, Sveits og Østerrike, der man ikke har screening (24).

Figur 1



Aldersspesifikk insidens av brystkreft i Sverige i perioden1980–84 og i perioden 1995–99. Screening foregår primært i aldersgruppen 50–69 år, og dette er merket med vertikale streker

Figur 2



Aldersspesifikk dødelighet av brystkreft i Sverige i perioden 1985–89 og 1995–99. Mammografiscreening i Sverige dekket 92 % av befolkningen i 1991. Screening foregår primært i aldersgruppen 50–69 år, og dette er merket med vertikale streker

Beregning av effektivitet

Det finnes flere metoder for å beregne effektiviteten av screening. En enkel metode er å telle antall prøver man må ta for å spare ett liv. Harris & Leininger (25) har regnet ut at man tar mellom 1 700 og 5 000 mammografier for hvert liv som blir spart. Hvis man skal evaluere screening med stor grad av overdiagnostisering, bør man justere for heterogenitet. Den enkleste måten er å sammenlikne stadiespesifikke rater før og etter start av screening. (Men det er bare i Norge man registrerer stadium ved diagnosetidspunkt.) Følgende to metoder kan brukes for å evaluere effekten og justere for heterogenitet:

Man kan sammenlikne antall overdiagnostiserte tilfeller med antall sparte liv. Hvis insidensen av brystkreft og carcinoma in situ til sammen øker med 70 % (10, 26) og forventet reduksjon i dødelighet er 20 %, som de to nyeste metaanalysene viser (11, 23) og 35 % av kreftpasientene normalt dør av sykdommen, er det ti friske kvinner som overdiagnostiseres for hvert liv som reddes.

Hvis man har historiske kreftrater, kan man bruke dette som forventet antall tilfeller og sammenlikne med antall tilfeller av intervallkreft for å beregne effektiviteten (1). Gjennomsnittlig antall krefttilfeller per år i fylkene som deltok i prøveprosjektet med mammografi var 300 før screening, og 75 % deltok i mammografiprogrammet (26). I perioden 1996–99 oppdaget man 247 tilfeller av brystkreft mellom første og andre screeningrunde (intervallet var på to år) (25). Forventet antall tilfeller som ville blitt oppdaget uten mammografi over to år er $300 \times 2 \times 0{,}75 = 450$. Det vil si at bare 55 % (247 av 450) av de krefttilfeller som man hadde før start av screening, blir oppdaget med mammografi.

Beregning av testintervall

Det finnes to fundamentale prinsipper for beregning av testintervall. Det første er at jo flere tester man tar, desto mer oppdager man. Hvis sensitiviteten er under 100 %, så vokser den kumulative sensitiviteten hvis man tar flere tester. Dessuten øker sannsynligheten for å oppdage rasktvoksende svulster ved å ta flere tester med kortere intervaller.

Det andre prinsippet er at man skal teste individer med høyere risiko oftere enn individer med lav risiko. Hvis man ikke vet noe om risikofaktorer, bør alle testes med lik hyppighet. Celleforandringer er vanligst hos kvinner mellom 25 og 50 år og henger sammen med seksuell aktivitet (27). Hvis

formålet er å oppdage flest mulig celleforandringer, bør disse kvinnene testes oftere enn kvinner over 50 år.

Opportunistisk screening versus organisert screening

Det er et medisinsk dogme at organisert screening alltid er mer effektivt enn opportunistisk screening. Dette dogmet baserer seg på sammenlikning av forekomst av livmorhalskreft i Norge, Finland, Sverige og England i perioden 1960–80 (28), hvor det ble observert en større reduksjon i Finland og Sverige enn i Norge og England. Denne forskjellen ble forklart med at organisert screening er mer effektivt enn opportunistisk screening. I en sammenlikning av 17 populasjoner fant Gustafsson og medarbeidere (27) ingen systematisk sammenheng mellom reduksjon i dødelighet og organisering av screening for livmorhalskreft. Disse resultatene er også vist i kvantitative analyser (29). En sammenlikning av livmorhalskreft i Sverige (med organisert Pap-screening) og Norge (med uorganisert Pap-screening) i perioden 1980–99 viste at aldersjustert insidens er uforandret i begge land (5, 16), men aldersjustert dødelighet sank med 30 % i både Sverige (5) og Norge (24).

Hva er det aktuelt å screene for?

Forebyggende behandling er morgendagens helsemarked. Svært mange diagnostiske prosedyrer kan brukes til screening. Sensitivitet, spesifisitet, pris og risiko ved bruk bestemmer anvendeligheten av screening – i tillegg til etikk, politikk og økonomi.

Litteratur

Komplett litteraturliste finnes i artikkelen på www.tidsskriftet.no

1. Armitage P, Colton T, red. Encyclopedia of biostatistics. Chichester: John Wiley & Sons, 1999: 3976–4022.
2. Myers ER, McCrory D, Nanda K, Bastian L, Matchar DB. Mathematical model for the natural history of human papillomavirus infection and cervical carcinogenesis. Am J Epidemiol 2000; 151: 1158–71.
3. Rothman KJ, Greenland S. Modern epidemiology. Philadelphia: Lippincott, Williams & Wilkins, 1998.
4. Cancer incidence in Sweden 1999. Stokholm: Socialstyrelsen, 2001.
5. Physician Insurers Association of America. Breast cancer study. Washington D.C.: Physician Insurers Association of America, 1995.
6. Http: //www.nih.gov/cancer_information/testing/
7. Kerlikowske K, Grady D, Barclay J, Sickles EA, Ernster V. Likelihood ratios for modern screening mammography. Risk of breast cancer based on age and mammographic interpretation. JAMA 1996; 276: 511–8.
8. Spratt JS, Meyer JS, Spratt JS. Rates of growth of human neoplasms: part II. J Surg Oncol 1996; 86: 68–83.
9. Klemi PJ, Joensuu H, Toikkanen S, Tuominen J, Rasanen O, Tyrkko J et al. Aggressiveness of breast cancers found with and without screening. BMJ 1992; 304: 467–9.
10. Tabár L, Fagerberg CJG, Gad A, Baldetorp L, Holmberg LH, Gröntoft O et al. Reduction in mortality from breast cancer after mass screening with mammography. Lancet 1985; 1: 829–32.