TREBALL DE FI DE MÀSTER

Màster en Estadística i Investigació Operativa

Títol: Correcció dels biaixos, relacionats amb la detecció precoç, de les funcions de supervivència del càncer de mama

Autor: Albert Roso Llorach

Directores: Montserrat Rué, Guadalupe Gómez

Departament: Estadística i Investigació Operativa

Convocatòria: Gener 2011

Facultat de Matemàtiques i Estadística

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Treball Fi de Màster

Correcció dels biaixos, relacionats amb la detecció precoç, de les funcions de supervivència del càncer de mama

Albert Roso Llorach

Directora: Montserrat Rué
Departament de Ciències Mèdiques Bàsiques, Universitat de Lleida

CoDirectora: Guadalupe Gómez
Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya

A la meua família, la gent de Lleida i de Tortosa

Resum

Des que es va començar a utilitzar la mamografia de cribratge per a diagnosticar precoçment el càncer de mama, les estimacions de supervivència de les dones que han patit càncer de mama estan afectades per dos biaixos. Un d'ells és el biaix que es produeix pel fet d'avançar el diagnòstic (*lead-time bias*). L'altre s'origina perquè els tumors detectats en els exàmens de cribratge tenen una probabilitat més alta de tenir un creixement més lent que els detectats en intervals entre exàmens o els d'un grup no cribrat (*length bias*).

Diversos autors han estudiat l'efecte dels dos biaixos esmentats en l'estimació de les funcions de supervivència en dones que participen en programes de cribratge. Alguns autors fan l'assumpció que el temps d'avenç del diagnòstic (*lead-time*) és independent del temps de supervivència posterior (*post-lead-time*). Altres autors, posen en dubte aquesta assumpció i modelitzen certa dependència entre els dos temps. L'objectiu ha estat obtenir estimacions correctes del temps de supervivència corregint pels biaixos esmentats.

S'han revisat diferents mètodes per a obtenir estimacions de la supervivència del càncer de mama lliures dels biaixos *lead-time bias* i *length bias*. S'han utilitzat dades dels registres de càncer de Girona i Tarragona i del programa de detecció precoç de l'Hospital del Mar. S'ha dut a terme un estudi de simulació per tal de reproduïr la història natural de la malaltia i estimar els paràmetres relacionats amb la detecció precoç.

S'han obtingut estimacions del temps d'avenç del diagnòstic en diferents escenaris de cribratge. S'han avaluat les diferències entre els casos detectats per examen i els casos de càncer d'interval. El temps mig d'avenç del diagnòstic en les dones diagnosticades de càncer de mama per examen de cribratge es troba al voltant dels 5 anys. Els temps de sojorn en estat pre-clínic dels casos detectats per examen són superiors als de la resta de dones incidents. Exceptuant un dels mètodes, la resta han proporcionat resultats similars.

S'han obtingut estimacions del temps de supervivència del càncer de mama, corregint els biaixos associats a la detecció precoç. Els resultats obtinguts mostren la importància de considerar aquest biaixos en l'avaluació correcta de la supervivència.

Paraules clau: supervivència; càncer de mama; detecció precoç; lead-time bias; length bias

Abstract

Since the introduction of screening mammography for the early detection of breast cancer, the survival estimates of the screen-detected cancer cases, are affected by two biases. One is the bias that occurs because of the advance of the diagnosis (lead-time). The other arises because the tumors detected on screening examinations are more likely to have slower growth than the cases detected in the intervals between examinations and other groups of cancer cases not detected by screening (length bias).

Differents authors have studied the effect of these biases in the survival functions of the screen-detected women. Some authors make the assumption that time gained by early detection (lead-time) is independent of survival time after diagnosis (post-lead-time). Other authors question this assumption and propose new models that involves dependence between the two times. The goal was to obtain correct estimates of survival time correcting these biases.

Different methodologies used to obtain estimates of the survival of breast cancer free of lead-time and length biases have been reviewed. Data from the cancer registries of Girona and Tarragona and from the Hospital del Mar early detection program have been used. A simulation study that reproduces the natural history of breast cancer and allows to estimate the parameters related to early detection was performed.

Lead time esimates under different screening scenarios have been obtained. Differences between screen-detected and interval cases have been assessed. The mean lead-time for screen-detected women is almost 5 years. Screen-detected cases have higher sojourn times in pre-clinical state than other cases. Except one of the methods, the rest have provided similar results.

Bias-corrected survival estimates have been obtained. The results show the relevance of considering these biases when estimating survival.

Keywords: Survival; Breast Cancer; Early Detection; Lead-time; Length Bias

Índex general

Capítol 1
Introducció

1. Motivació

Aquest Treball Final de Màster s'integra en el projecte d'investigació del grup format per investigadors de la Facultat de Medicina de la Universitat de Lleida (UdL), de l'Institut d'Investigació Biomèdica de Lleida (IRBLleida), de la Universitat Rovira i Virgili (URV), de l'Institut Municipal d'Investigació Mèdica de Barcelona (IMIM) i de l'Hospital de Tortosa Verge de la Cinta (HTVC). El projecte d'investigació que el grup té en marxa actualment tracta de l'optimització dels programes de cribratge poblacional mitjançant models matemàtics (PS09/01340, IP: Montserrat Rué).

L'objectiu general del projecte és valorar els beneficis, efectes adversos i costos de diferents estratègies de detecció precoç del càncer de mama i comparar les alternatives mitjançant una anàlisi de cost-efectivitat. A més a més, es treballarà en l'optimització dels programes de cribratge poblacionals mitjançant l'adaptació del patró de cribratge (edats d'inici i finalització i interval entre exàmens) al risc de desenvolupar càncer de mama. Es pretén que els resultats obtinguts puguin ajudar en la presa de decisions de política sanitària.

2. Situació actual

La mortalitat per càncer de mama en els països occidentals ha seguit una tendència decreixent des de principis dels anys 1990 [**1**]. S'ha estimat que la utilització de la mamografia de cribratge i els tractaments adjuvants del càncer de mama han tingut un efecte similar en la millora de la supervivència del càncer de mama [**2**]. No obstant, el càncer de mama segueix sent el càncer més freqüent en les dones de tot el món i la primera causa de mortalitat prematura en les dones de 35 a 64 anys.

El cribratge en medicina és una estratègia utilitzada en una població per a detectar una malaltia en els individus sense signes o símptomes d'aquesta malaltia. A diferència de les intervencions mèdiques terapèutiques, en el cribratge es duen a terme intervencions en els individus que no presenten signes o símptomes de la malaltia i molts d'ells mai la patiran.

L'objectiu dels programes de cribratge en càncer de mama és reduïr la mortalitat per aquesta causa. Els exàmens de detecció permeten diagnosticar els càncers en una fase més precoç. Els mètodes usats ens les últimes dècades per a detectar el càncer de mama en una fase més precoç són: la mamografia, l'autoexamen i l'examen clínic de la mama. Amb el càncer detectat en una etapa inicial, el tractament pot ser més fàcil i eficaç, augmentant el temps de vida de la dona.

3. Biaixos de la detecció precoç

La reducció en mortalitat per càncer de mama és la forma acceptada de mesurar l'efectivitat dels exàmens selectius de detecció. Quan parlem de la detecció, avaluar les millores en supervivència des del moment del diagnòstic no és representatiu del benefici, ja que existeixen diversos biaixos únics de la detecció precoç que confonen el benefici [**3**].

3.1. Biaix d'avenç del diagnòstic (*lead-time bias*). Normalment la supervivència específica al càncer de mama es mesura a partir del moment en què es realitza el diagnòstic fins al moment de la mort. Si un càncer de mama és detectat per exàmens abans que es presentin símptomes, llavors el temps d'avenç per diagnòstic és igual al període entre la detecció per exàmens i el moment on s'haurien presentat els primers símptomes o signes.

Encara que el tractament precoç no hagi tingut cap benefici, la supervivència de les persones que han estat diagnosticades en exàmens de detecció és més llarga simplement perquè s'hi ha afegit el temps d'avenç (Figura 1.1). Per aquest fet, no es pot avaluar correctament el benefici de la detecció precoç a partir de la comparació dels temps de supervivència.

3.2. Biaix de durada de la malaltia en estat pre-clínic (*length bias*). Quan parlem del *length bias* ens referim a la probabilitat més elevada de detectar càncers menys agressius i de creixement més lent, quan s'utilitza una prova de cribratge (Figura 1.2). Els pacients amb càncers detectats per examen sobreviuen més temps en part perquè els càncers són menys agressius, per tant la millor supervivència no es pot atribuïr únicament al tractament precoç.

FIGURA 1.1. Biaix d'avenç del diagnòstic

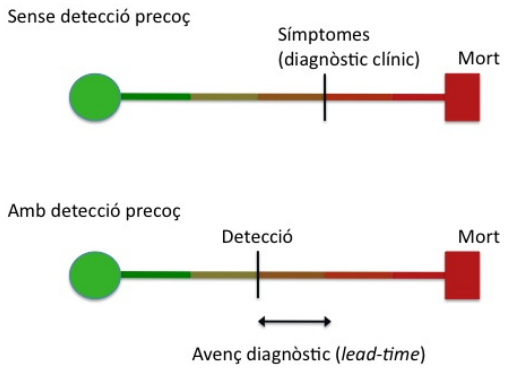
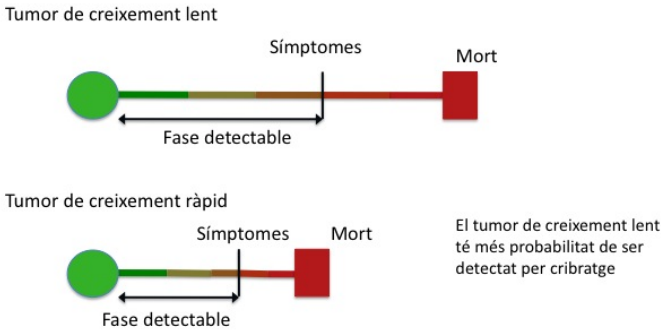


FIGURA 1.2. Biaix de durada de la malaltia en estat pre-clínic



4. Objectiu

L'objectiu principal del projecte és obtenir estimacions correctes del temps de supervivència del càncer de mama, controlant l'efecte dels biaixos *lead-time bias* i *length bias*.

Els resultats d'aquest projecte s'utilitzaran per avaluar l'impacte de diferents estratègies de detecció precoç, en el projecte de recerca esmentat en l'apartat 1.

5. Organització de la memòria

La memòria s'organitza de la següent manera:

En el capítol 2 s'explica breument la modelització de la detecció precoç del càncer de mama, es descriuen els biaixos *lead-time bias* i *length bias*. En el capítol 3 es presenten les diferents metodologies per a corregir els biaixos del cribratge. En el capítol 4 es presenta la metodologia utilitzada per analitzar les dades de supervivència dels registres poblacionals de càncer de Girona i Tarragona i del programa de detecció precoç de l'Hospital del Mar. Es descriu el disseny d'un estudi de simulació per a obtenir temps de durada de les fases pre-clínica i clínica del càncer de mama. En el capítol 5 es mostren els resultats obtinguts amb l'aplicació de les diferents metodologies a les dades de supervivència de càncer de mama reals i simulades. En el capítol 6 es discuteixen els resultats principals, les limitacions del projecte i els aspectes que podrien ser millorats en un futur.

Capítol 2
Antecedents i revisió de la literatura

Diferents models matemàtics han estat desenvolupats per a tractar les diferents situacions sorgides a partir dels programes de cribratge [**3, 4**]. Les primeres investigacions dels models es van dur a terme durant els anys 60, posteriorment s'han desenvolupat diferents mètodes estadístics per a l'avaluació dels calendaris d'exàmens. Alguns autors s'han centrat en discutir l'optimització dels programes de cribratge, tenint en compte l'edat d'inici i el temps entre exàmens[**5, 6**].

Un altre aspecte en l'estudi analític per a determinar els calendaris dels programes de cribratge és l'estimació dels paràmetres usats en els models estadístics. Aquests paràmetres són el temps mig de sojorn en l'estat pre-clínic i la sensibilitat dels exàmens de detecció. Aquestes estimacions han estat obtingudes per diferents autors a partir dels resultats de diferents assaigs clínics[**7, 8**].

1. Història natural de la malaltia

Seguint el marc teòric de Zelen, es considera una població on en qualsevol punt del temps un individu pot estar en un dels tres possibles estats [**3**]:

- S_0**:** Estat lliure de malaltia. L'individu no té la malaltia o la té en estat indetectable per la prova diagnòstica estudiada.
- S_p**:** Estat pre-clínic. La malaltia pot ser detectada mitjançant una prova diagnòstica, per exemple, la mamografia.
- S_c**:** Estat clínic. Han aparegut els símptomes i la malaltia ha estat diagnosticada pel procediment mèdic habitual.

S'assumeix que la història natural del càncer de mama correspon a la d'una malaltia progressiva amb transicions $S_0 \to S_p \to S_c$. A vegades és avantatjós afegir un estat

absorbent S_d que es refereixi a la mort de l'individu. Teòricament és possible entrar en l'estat absorbent des de qualsevol estat. Els camins habituals seran $S_c \to S_d$ o $S_0 \to S_d$. Però si el camí és $S_p \to S_d$ llavors implica que la persona mor sense que la malaltia hagi estat detectada o que ha estat diagnosticada sense benefici (sobrediagnòstic).

La Figura 2.1 descriu un cas típic d'història natural del càncer de mama. Representa l'edat de la dona en anys al llarg del temps. L'edat de la dona a l'origen de l'observació és z. Si es consideren els temps cronològics

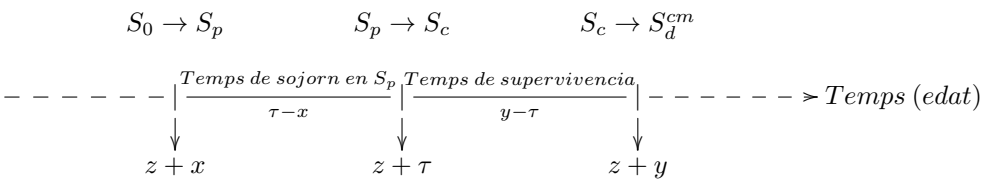
- x temps des de l'origen fins l'entrada a S_p.
- τ temps des de l'origen fins l'entrada a S_c
- y temps des de l'origen fins la mort

s'obtenen les edats

- $z + x$ edat a l'entrada a S_p
- $z + \tau$ edat a l'entrada a S_c
- $z + y$ edat de la mort

La diferència entre el temps d'entrada a pre-clínic i temps d'entrada a l'estat clínic correspon al temps de sojorn en l'estat pre-clínic $\tau - x$. Per altra banda, el temps de supervivència serà la diferència entre el temps en morir i el temps d'entrada a l'estat clínic $y - \tau$.

FIGURA 2.1. Estats i transicions en el càncer de mama



2. Biaix d'avenç del diagnòstic (*lead-time bias*)

Suposi's que una població participa en un programa de cribratge poblacional. L'ideal seria que els membres de la població que pateixen la malaltia pre-clínica fossin detectats per examen. Si aquests individus no haguessin estat sotmesos a examen, la malaltia hauria estat detectada més tard, quan hagués progressat a l'estat clínic. El temps pel qual s'avança el diagnòstic mitjançat examen s'anomena *lead-time*. En el llenguatge de la teoria de la probabilitat el *lead-time* és sinònim de *forward recurrence time* [**3**].

Suposi's que una persona és detectada estant a l'estat pre-clínic al temps t_0. Quant de temps s'hagués quedat a S_p fins que hagués entrat a S_c? Aquest temps és el *lead-time* o *forward recurrence time*.

Noti's que la probabilitat d'un individu d'estar a S_p depèn de la probabilitat de transició d'S_0 a S_p i del seu temps de sojorn en estat pre-clínic. L'origen del temps i el punt t_0 es consideren temps cronològics. Des del moment que l'origen del temps és arbitrari, es pot considerar t_0 tan lluny com vulguem de l'origen.

Serà convenient adoptar la següent notació:

$q(t)$ és la funció de densitat de probabilitat de T, definida com la variable aleatòria del temps de sojorn en S_p;

$Q(t) = \int_t^\infty q(x)\,dx$ és la funció de supervivència de T

$P(t)$ és la probabilitat d'estar a S_p al temps t;

$w(t)\,dt$ és la probabilitat de transició de S_0 a S_p durant $(t, t+dt)$;

$Q_f(t|t_0)$ és la probabilitat condicionada de que un individu que és a S_p al temps t_0 es quedi en aquest estat almenys t unitats addicionals de temps;

$q_f(t|t_0) = -\frac{d}{dt}Q_f(t|t_0)$ és la funció de densitat de probabilitat del *forward recurrence time*.

La probabilitat que una persona estigui a S_p al temps t_0 i es quedi en aquest estat almenys t unitats addicionals de temps és $P(t_0)Q_f(t|t_0)$. Aquest esdeveniment pot succeïr si una persona ha entrat en S_p durant el temps $(t_0 - x, t_0 - x + dx)$ i s'hi ha quedat almenys $(t + x)$ unitats de temps. Per tant, podem escriure:

$$P(t_0)Q_f(t|t_0) = \int_0^{t_0} w(t_0 - x)Q(t + x)\,dx. \tag{1}$$

Aquesta expressió permet calcular la distribució no condicionada del *forward recurrence time*, tal com descriu Zelen [**3**]. Aquesta distribució està definida per

$$Q_f(t) \quad = \frac{\int_t^\infty Q(y)\,dy}{m}, \tag{2}$$

on

$$m = \int_0^\infty Q(x)\,dx$$

és l'esperança del temps de sojorn en S_p.

Suposi's que un programa de detecció precoç on la distribució del *forward recurrence time* té funció de densitat

$$q_f(t) = \begin{cases} Q(t)/m & t \geq 0 \\ 0 & t < 0 \end{cases}$$

Llavors el primer moment de la distribució del *forward recurrence time* és el *lead-time* esperat. Si L indica el primer moment, aleshores

$$L = \int_0^\infty t q_f(t)\, dt = \frac{m^2 + \sigma^2}{2m} = \frac{m}{2}(1 + C^2), \tag{3}$$

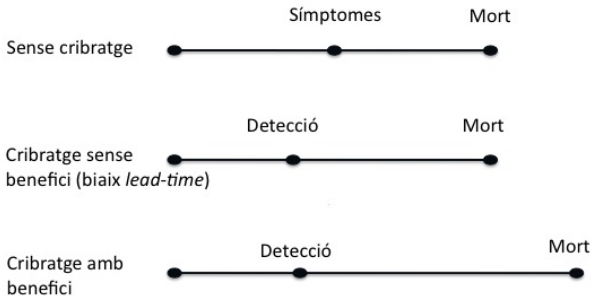
on $C = \sigma/m$, i m i σ^2 són la mitjana i la variància del temps de sojorn en S_p. L és el temps mitjà guanyat amb el diagnòstic precoç mitjançant l'examen de cribratge. Si $q(t)$ segueix una distribució exponencial de mitjana m, llavors $C = 1$ i el *forward recurrence time* té la mateixa distribució exponencial que $q(t)$.

Noti's que $L > \frac{1}{2}m$ per a $\sigma^2 > 0$. Si el punt on es fa l'examen de cribratge és seleccionat aleatòriament respecte al temps de sojorn en S_p, hom pot creure que el *lead-time* esperat serà $\frac{1}{2}m$. Però cal tenir en compte que la distribució de probabilitat del temps de sojorn per a tots aquells individus detectats per cribratge és diferent de la funció de densitat de probabilitat de la població general $q(t)$. El cribratge no detecta individus aleatòriament sinó que detecta als individus amb temps de sojorn en estat pre-clínic més llargs. Aquest fenomen s'anomena mostreig afectat pel biaix de durada (*length biased sampling*).

Cal remarcar que s'ha de tenir especial cura si es vol comparar la supervivència dels individus detectats precoçment amb la supervivència d'un grup control que conté individus que han estat diagnosticats clínicament. Si el mètode de detecció precoç detecta la malaltia abans que el mètode habitual, i no hi ha cap millora deguda al tractament, la persona no obtindrà cap benefici. Amb els casos detectats precoçment, la supervivència semblarà més llarga encara que no hi hagi cap benefici. A la figura 2.2 es descriu aquesta situació.

L'inici del seguiment (moment del diagnòstic) de la supervivència dels individus detectats mitjançant cribratge està situat en l'estat pre-clínic, mentre que l'inici del grup control és en l'estat clínic. En conseqüència, per tal de realitzar una comparació, les dades de supervivència dels dos grups han de tenir el mateix punt de partida. Això suposa que el temps mig de supervivència del grup detectat en estat pre-clínic hauria de ser corregit restant-li el *lead-time* esperat.

FIGURA 2.2. Il·lustració del biaix *lead-time*

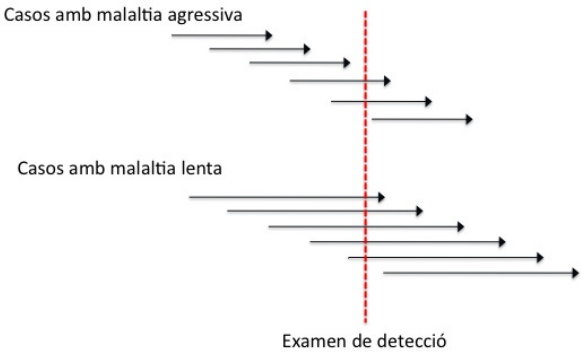


3. Biaix de durada (*length bias*)

El biaix de durada o *length bias* sorgeix quan els individus detectats per cribratge no conformen una mostra aleatòria de la població que es troba en estat pre-clínic. Els individus detectats en S_p tendeixen a tenir uns temps de sojorn en estat pre-clínic més llargs. Hi ha evidències biològiques que mostren una correlació de la fase clínica amb la fase pre-clínica. Un temps de sojorn curt en pre-clínic implica que la malaltia és agressiva mentre que un llarg temps de sojorn implica que la malaltia té un creixement lent. Per tant, els individus amb llargs temps de sojorn en estat pre-clínic tenen més possibilitats de viure més temps en comparació als que tenen una durada pre-clínica curta. Per conseqüència, des del moment en què l'examen de cribratge té més probabilitat de diagnosticar individus amb temps de sojorn pre-clínics més llargs, aquest grup tendirà a tenir temps de supervivència més llargs.

Es considera una població de gent amb la malaltia en estat pre-clínic com a la Figura 2.3. D'acord amb el model exposat la longitud de les línies horitzontals corresponen a la durada del temps de sojorn en estat pre-clínic. El temps de sojorn en estat pre-clínic diferirà entre els individus i seguirà una distribució de probabilitat. Es defineix la sensibilitat d'un examen com la probabilitat d'un resultat positiu per a un individu cribrat durant la fase pre-clínica. Si l'examen de detecció precoç té sensibilitat igual a 1, la detecció de casos és equivalent a plantar una línia vertical en un punt aleatori del temps. La intersecció de la línia vertical amb la

FIGURA 2.3. Il·lustració del biaix de durada (*length bias*)



línia horitzontal correspon a un cas diagnosticat. Clarament la línia vertical té més probabilitats d'intersecar-se amb una línia horitzontal llarga que amb una de curta, i en resulta una mostra afectada pel biaix de durada (*length biased sample*). El fenomen del biaix de durada succeix amb independència del valor de la sensibilitat de l'examen.

4. Mostreig afectat pel biaix de durada (*length biased sampling*)

Els casos diagnosticats per examen no són diagnosticats aleatòriament com s'ha vist a la secció 3, sinó que constitueixen una mostra afectada pel biaix de durada. Com més temps estigui l'individu en l'estat pre-clínic més gran serà la probabilitat de ser detectat en un examen de cribratge. Si el curs clínic de la malaltia està positivament correlacionat amb el curs pre-clínic, llavors els individus diagnosticats en un programa de detecció precoç és probable que visquin més temps perquè la malaltia és de creixement lent. Com a resultat, la supervivència és un 'endpoint' inapropiat per tal d'avaluar el benefici d'un programa de detecció precoç. En aquesta secció s'il·lustren les principals propietats estadístiques del mostreig afectat pel biaix de durada [**9**].

El mostreig afectat pel biaix de durada assumeix que la probabilitat de diagnosticar un cas en estat pre-clínic és proporcional al temps de sojorn en S_p. Es pot representar la probabilitat de diagnosticar un cas a S_p definint T com la variable aleatòria del temps de sojorn en S_p amb funció de densitat de probabilitat $q(t)$ i

$$a_0 = \begin{cases} 1 & \text{si és diagnosticat a } S_p \\ 0 & \text{altrament} \end{cases}$$

Per una valor concret $T = t$, la probabilitat de diagnosticar un cas a S_p és

$$P\{a_0 = 1 | t < T \leq t + dt\} \propto t.$$

Llavors la probabilitat conjunta és

$$P\{a_0 = 1, t < T \leq t + dt\} \propto tq(t)dt$$

resultant en

$$\begin{aligned} P\{t < T \leq t + dt | a_0 = 1\} &= tq(t)dt/m \\ f(t|a_0 = 1) &= tq(t)/m \end{aligned} \tag{4}$$

on

$$m = \int_0^\infty tq(t)\, dt.$$

La mitjana del temps de sojorn en S_p condicionada a ser diagnosticat precoçment és

$$E(T|a_0 = 1) = E(T^2)/m = m(1 + C^2),$$

on $C = \sigma/m$, el que demostra que aquest temps de sojorn és més llarg que el temps de sojorn de la població general. En el cas que T sigui exponencial, $C = 1$ i $E(T|a_0 = 1) = 2m$.

Capítol 3
Mètodes seleccionats de la bibliografia

1. Mètode de Walter & Stitt

1.1. Resum. La metodologia presentada per Walter & Stitt [**10**] descriu la supervivència dels casos detectats per cribratge mitjançant una funció de risc, que en general depèn de la duració del temps de sojorn en estat pre-clínic (que segueix una distribució exponencial de paràmetre λ), el *lead-time* i la supervivència total des del diagnòstic, que es compon de la suma de dues components: el *lead-time* i el *post-lead-time* (temps des d'on hauria succeït el diagnòstic clínic en absència de cribratge fins a la mort, etc.). Els autors estudien dues versions de la funció de risc, una on assumeixen un risc constant i una altra on consideren que el risc depèn del *lead-time* i la durada del temps de sojorn. A partir de l'estimació màxim verssemblant del *post-lead-time*, es pot comparar la supervivència dels casos detectats per cribratge amb la dels no detectats per cribratge (casos d'interval, individus no cribrats, etc.)

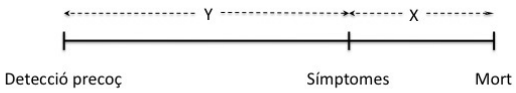
1.2. Model. En el mètode presentat, els autors assumeixen els models de detecció de malalties cròniques descrits en anteriors treballs del seu grup [**8**]. La història natural de la malaltia correspon a la d'una malaltia progressiva amb un estat lliure de malaltia, un estat pre-clínic i un estat clínic. Es defineix el *lead-time* com una funció de la distribució del temps de sojorn en estat pre-clínic, la sensibilitat de la prova de cribratge, i el calendari d'exàmens aplicats a la població. L'anàlisi de la distribució del temps de sojorn es va fer amb les dades provinents de l'estudi HIP (Health Insurance Plan of Greater New York), un assaig aleatori dissenyat per a l'avaluació de l'efecte dels programes de cribratge i la reducció de la mortalitat del càncer de mama entre les dones de 40-64 anys. La distribució que va obtenir un millor ajust a les dades va ser la distribució exponencial [**7**].

En aquest model, el temps de supervivència per a tots els casos detectats per cribratge s'expressa a partir de dues components:

- el *lead-time*, consisteix en el temps des del diagnòstic fins al temps on hagués succeït la detecció clínica en absència de cribratge.
- el temps de supervivència després des del moment del diagnòstic, fins a la mort.

Esquemàticament

FIGURA 3.1. Esquema supervivència



A la figura 3.1 Y denota el *lead-time*, i X el temps de supervivència 'extra' després del moment esperat de la detecció clínica. L'objectiu de l'anàlisi presentat és l'estimació de la distribució d'X per als casos detectats per cribratge; llavors es compara aquesta distribució amb la supervivència de casos no detectats per cribratge, per a examinar si hi ha hagut millora.

El model assumeix que els individus detectats per cribratge sobreviuen com a mínim la durada del *lead-time*. Encara que existeixen efectes adversos del cribratge (sobrediagnòstic, tractaments agressius,..) que podrien incrementar el risc de mort, els autors consideren que aquestes situacions són excepcionals i generalment el cribratge comporta un benefici en la supervivència. Per tant, es pot considerar $X \geq 0$.

La supervivència dels casos detectats per cribratge és descrita mitjançant una funció de risc, que depèn de la durada del temps de sojorn T, el *lead-time* Y, i el temps de supervivència total des del diagnòstic Z. Denoten aquesta funció com $h(t, y, z)$. A partir de la relació entre la supervivència i el risc, la distribució condicional del *post-lead-time* per a un pacient determinat amb valors $Y = y$ i $T = t$ és

$$p(x|y,t) = h(t,y,x+y)\exp\left[-\int_0^{x+y} h(t,y,u)\,du\right] \tag{5}$$

Per tant, la distribució marginal d'X serà

$$p(x) = \int_0^{\infty}\int_0^{t} h(t,y,x+y)\exp\left[-\int_0^{x+y} h(t,y,u)\,du\right] g(y|t)q(t)\,dydt \tag{6}$$

on $g(y|t)$ és la distribució condicional del *lead-time*, per a un temps t determinat, i $q(t)$ és la funció de densitat del temps de sojorn T.

Segons els autors, en el cas que el temps de supervivència 'extra' X fos observable, es podria comparar amb el temps de supervivència dels casos no cribrats. Per a un pacient determinat, només Z seria observable mentre que les variables X, Y i T són desconegudes. Per tant, s'hauran d'estimar les distribucions d'X i Y a partir de dades observades de supervivència des del diagnòstic dels casos detectats per cribratge, per tal de realitzar inferència sobre les dues components del model.

1.2.1. *Model senzill.* En primer lloc, els autors presenten una modelització simple per a la funció de risc. En aquest model assumeixen que el risc és independent del *lead-time* i la durada del temps de sojorn en estat pre-clínic. La funció de risc té la forma

$$h(t,y,z) = \begin{cases} 0 & z \leq y \\ \theta & z > y \end{cases} \tag{7}$$

Amb aquesta funció de risc assumeixen un temps de garantia des del moment de la detecció precoç fins al moment en què la malaltia s'hagués tornat clínica, on no existeix cap risc per a l'individu. A partir de la detecció clínica el risc és constant amb valor θ. Sota aquest model i aplicant l'equació 6, la distribució condicional de Z, donat $Y = y$, que obtenen és

$$p(z|y) = \begin{cases} 0 & z \leq y \\ \theta \exp\left[-\theta(z-y)\right] & z > y \end{cases}$$

Per a obtenir la distribució no condicionada de Z, calculen la funció de distribució conjunta de Z i Y i integren sobre el suport d'Y. A partir dels resultats de l'estudi HIP, obtenen que la distribució exponencial ajusta millor el temps de sojorn en estat pre-clínic [**7**]. Llavors poden assumir que el *lead-time* Y és distribueix exponencialment amb paràmetre λ. Si el suport d'Y ve determinat per la condició $Y \leq Z$, llavors la distribució de Z és

$$\begin{aligned} p(z) &= \int_0^z \theta \exp\left[-\theta(z-y)\right] \lambda \exp\left[-\lambda y\right] \, dy \\ &= \begin{cases} \frac{\lambda\theta}{\lambda-\theta}\left[\exp(-\theta z) - \exp(-\lambda z)\right] & \lambda \neq \theta \\ \lambda^2 z \left[\exp(-\lambda z)\right] & \lambda = \theta \end{cases} \end{aligned} \tag{8}$$

Els autors, per problemes d'identificabilitat en l'estimació dels paràmetres causats per la simetria de la expressió en λ i θ, condicionen λ a les dades de prevalença i incidència de la població cribrada, i després la utilitzen en l'estimació màxim versemblant per a obtenir el paràmetre θ.

Les dades de supervivència estan agrupades en intervals de temps. Per tal d'ajustar les dades agrupades, es necessita la probabilitat de morir durant l'interval de temps, condicionada a què la persona és viva a l'inici del interval. A partir de 8 calculen la funció de probabilitat acumulada quan $\lambda \neq \theta$

$$P(z_i) = 1 + \frac{\theta\left[\exp(-\lambda z_i)\right] - \lambda\left[\exp(-\theta z_i)\right]}{\lambda - \theta}$$

Llavors, per a l'interval (z_i, z_{i+1}),

$$P(z_i, z_{i+1}) = \frac{\theta\left[\exp(-\lambda z_{i+1}) - \exp(-\lambda z_i)\right] - \lambda\left[\exp(-\theta z_{i+1}) - \exp(-\theta z_i)\right]}{\lambda\left[\exp(-\theta z_i)\right] - \theta\left[\exp(-\lambda z_i)\right]} \tag{9}$$

Les dades de taula de vida consisteixen en el nombre de persones n_i a risc i el nombre de morts d_i durant l'interval (z_i, z_{i+1}). La versemblança combinada de les dades resulta en un producte de contribucions binomials $P(z_i, z_{i+1})$, una per a cada interval. La funció de log-versemblança és de la forma

$$logL = \sum_i d_i P(z_i, z_{i+1}) + (n_i - d_i)(1 - P(z_i, z_{i+1}))$$

Amb l'estimació màxim versemblant, aconsegueixen estimacions de θ i λ, i infereixen la distribució d'X. A partir del model 7, que assumia que el risc era independent del *lead-time* i del temps de sojorn, obtenen que la distribució del *post-lead-time* x segueix una exponencial de paràmetre θ. Amb aquests resultats, el autors consideren que el *lead-time* Y i el *post-lead-time* X són independents i es distribueixen exponencialment amb diferents paràmetres.

1.2.2. *Model alternatiu.* El model anterior 7, tot i la seva simplicitat, és en general poc realista. Sobretot l'assumpció d'un risc independent del *lead-time* i del temps de sojorn. Per tal de caracteritzar aquesta relació, els autors proposen un model més general per a la funció de risc

$$h(t, y, z) = \begin{cases} 0 & z \leq y \\ \theta & z > y \text{ i } y < Y^* \\ k\theta & z > y \text{ i } y \geq Y^* \end{cases} \tag{10}$$

En aquest model assumeixen que existeix un *lead-time* crític Y^*; si la detecció per examen produeix un *lead-time* més gran o igual que Y^* llavors el risc és modificat per un factor k. Si el *lead-time* és menor que Y^* el risc és manté constant a θ, i durant el temps en la fase pre-clínica el risc és nul.

A partir d'aquest model, consideren que un *lead-time* llarg correspon a un millor pronòstic, llavors esperen que $k < 1$. Si k és propera a 1, l'escenari és el que descriuen els autors amb el model anterior 7. Seguint el mateix raonament obtenen la distribució de Z si $\lambda \neq \theta$

(11)
$$p(z) = \begin{cases} \frac{\theta\lambda}{\lambda-\theta}\left[e^{-\theta z} - e^{-\lambda z}\right] & z \leq Y^* \\ \frac{\theta\lambda}{\lambda-\theta}e^{-\theta z}\left[1 - e^{(\theta-\lambda)Y^*}\right] + \frac{k\theta\lambda}{\lambda-k\theta}e^{-k\theta z}\left[e^{(k\theta-\lambda)Y^*} - e^{(k\theta-\lambda)z}\right] & z > Y^* \end{cases}$$

Al seu torn, si se suposa que la distribució del *lead-time* és exponencial de paràmetre λ [**7**], la distribució condicional d'X donat Y és

$$p(x|y) = \begin{cases} \theta e^{-\theta x} & y \leq Y^* \\ k\theta e^{-k\theta x} & y > Y^* \end{cases}$$

Llavors poden derivar la distribució d'X

$$p(x) = \theta e^{-\theta x} - e^{-\theta Y^*}\left[\theta e^{-\theta x} - k\theta e^{-k\theta x}\right]. \tag{12}$$

Els casos on $k < 1$ i $k > 1$ corresponen a una correlació positiva o negativa del *lead-time* i el *post-lead-time*. En general, quan s'han d'ajustar les dades empíriques cal estimar els paràmetres θ, k i Y^*.

Els autors proposen un segon model alternatiu tenint en compte l'efecte del *length bias* en la supervivència. Suposen que existeix un valor crític T^* per al temps de sojorn, amb la corresponent funció de risc

$$h(t, y, z) = \begin{cases} 0 & z \leq y \\ \theta & z > y \text{ i } t < T^* \\ f\theta & z > y \text{ i } t \geq T^* \end{cases} \tag{13}$$

Si $f < 1$, els casos amb temps de sojorn en estat pre-clínic llargs tenen menys risc que els casos amb temps curts. Desenvolupant el model per a calcular les distribucions de probabilitat, els autors obtenen que els models 10 i 13 són equivalents però amb paràmetres diferents. La interpretació que en donen els autors és que sota les hipòtesis del model és impossible diferenciar si el risc per als casos detectats per cribratge depèn del *lead-time* i/o del temps de sojorn en estat pre-clínic.

2. Mètode de Xu & Prorok

2.1. Resum. Xu i Prorok [**11**] estenen el model de Walter & Stitt [**10**], relaxant l'assumpció paramètrica per al *post-lead-time*. Es considera el temps total de supervivència com la suma del *lead-time* i del *post-lead-time* sota la hipòtesi d'independència i que aquest *lead-time* es distribueix exponencialment, assumint que la seva distribució és totalment coneguda. Les estimacions del *post-lead-time*

s'obtenen a partir de l'estimació màxim versemblant no paramètrica. L'estimador d'aquest nou model aconsegueix un millor ajust a les dades de supervivència estudiades per Walter & Stitt [**10**].

2.2. Model. El model assumeix que la història natural de la malaltia correspon a la d'una malaltia progressiva amb transicions $S_0 \to S_p \to S_c$. Es defineixen les variables Y com el *lead-time* i X com el temps de supervivència extra, o *post-lead-time*. El temps total de supervivència des del diagnòstic per cribratge Z ve determinat per l'expressió

$$Z \stackrel{d}{=} X + Y, \tag{14}$$

on d significa que les variables aleatòries de les dues bandes de la igualtat tenen la mateixa distribució. Com el punt del diagnòstic clínic ($S_p \to S_c$) d'un individu detectat per cribratge és generalment desconegut, no és possible observar les variables X i Y. Només es disposa d'informació de Z per a tots aquests individus. L'objectiu dels autors és poder estimar la funció de supervivència $\bar{F}_X(x) = P(X > x)$ d'X basada en observacions de Z.

Els autors fan certes assumpcions sobre X i Y per a obtenir el model 14, també anomenat model de deconvolució. L'assumpció principal és que la distribució del *lead-time* Y és completament coneguda, i la seva funció característica és diferent de zero. Per altra banda, aquest *lead-time* Y ha de ser no negatiu per a la població cribrada. Els autors no tenen en compte el cas en què el tractament posterior a la detecció per cribratge comporti un risc de mort més gran que el cas on l'individu no hagués estat diagnosticat. Per tant, el temps de supervivència extra X ha de ser no negatiu. Llavors, assumeixen que X i Y són independents i la distribució d'Y és exponencial amb paràmetre λ conegut.

Amb aquestes condicions els autors obtenen la distribució d'X del model 14, en termes de Z i Y. Noti's que $G(Y) \stackrel{d}{=} 1 - G(Y)$ segueix una distribució uniforme en l'interval $(0, 1)$ atès que G és una funció de distribució. Amb aquesta propietat, realitzen les següents transformacions:

$$\begin{aligned} W &= 1 - G(Z) = \exp(-\lambda Z), \\ U &= 1 - G(Y) = \exp(-\lambda Y), \\ V &= 1 - G(X) = \exp(-\lambda X). \end{aligned} \tag{15}$$

Sota aquestes condicions, el model 14 és equivalent a

$$W \stackrel{d}{=} UV, \tag{16}$$

on U i V són independents i U segueix una distribució uniforme sobre l'interval $(0,1)$. Les transformacions 15 i el model 16 permeten l'estimació d'una funció de densitat monòtona decreixent [**12**].

Els autors denoten per F_Z, F_V i F_W les funcions de distribució de Z, V i W respectivament, i $\bar{F}_i = 1 - F_i$, on $i = V, W, Z$, com les funciones de supervivència. A partir de l'expressió 16 obtenen

$$f_W(w) = \frac{d}{dw} F_W(w) = \int_w^\infty \frac{1}{v} f_V(v)\, dv = \int_w^\infty \frac{1}{v} dF_V(v). \tag{17}$$

Resolent la integral poden obtenir $\bar{F}_V(w)$ en termes de F_W i f_W:

$$\bar{F}_V(w) = \bar{F}_W(w) + w f_W(w). \tag{18}$$

Per altra banda, calculant directament de 15

$$\begin{aligned} F_X(x) &= P(X \leq x) = \bar{F}_V(e^{-\lambda x}), \\ F_Z(z) &= P(Z \leq z) = \bar{F}_W(e^{-\lambda z}). \end{aligned} \tag{19}$$

Diferenciant F_Z en 19 respecte a z i utilitzant 18 i 19 obtenen

$$\bar{F}_X(z) = \bar{F}_Z(z) - \frac{f_Z(z)}{\lambda}, \tag{20}$$

on $f_Z(z)$ és la funció de densitat de Z. Així, sota els supòsits del model, a 20 obtenen un mètode per a corregir o ajustar el biaix *lead-time* de la supervivència dels casos detectats per cribratge. D'aquesta equació es pot estimar $\bar{F}_X(z)$ si es pot estimar $\bar{F}_Z(z)$ i $f_Z(z)$. Com que la funció de densitat $f_W(w)$ de W és sempre monòtona decreixent, els estimadors màxim versemblants no paramètrics (NPMLE) de $F_W(w)$ i $f_W(w)$, basats en dades de $W_1, ..., W_n$ de $F_W(w)$ són, respectivament

$F^*_{W_n}(w) =$ menor majorant còncava de la funció de distribució empírica $F_{W_n}(w)$ de $W_1, ..., W_n$,

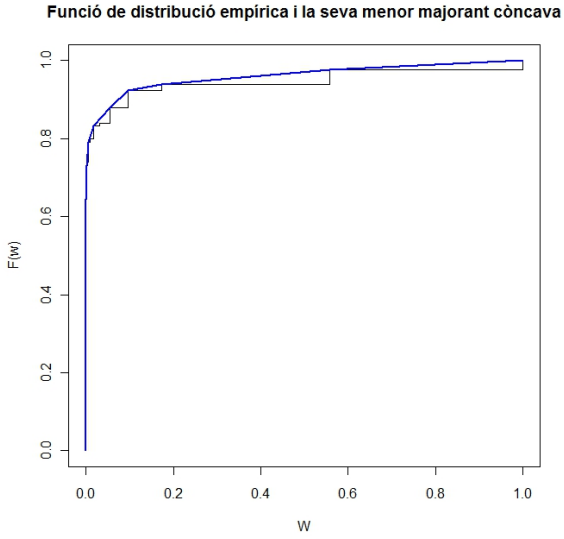
$$f^*_{W_n}(w) = \lim_{h \to 0^+} \frac{F^*_{W_n}(w) - F^*_{W_n}(w-h)}{h},\ w < W_{(n)}$$

$W_{(n)} = \max(W_i, ..., W_n)$.

Per tant, $f^*_{W_n}(w)$ és el pendent de la menor majorant còncava de $F_{W_n}(w)$ a w [**12**]. En general, no hi ha una fórmula explícita per a expressar $F^*_{W_n}(w)$ ja que depèn de les dades. Es pot obtenir $F^*_{W_n}(w)$ utilitzant línies rectes per a connectar punts

de salt apropiats de la funció de distribució empírica $F_{W_n}(w)$ de tal manera que $F^*_{W_n}(w)$ és la funció de distribució menor concàva respecte a $F_{W_n}(w)$ (Figura 3.2).

FIGURA 3.2. Funció de distribució empírica i la seva menor majorant còncava



Utilitzant la relació 19, es pot obtenir el NPMLE $\bar{F}^*_{Z_n}$ de $\bar{F}_Z$, definit per

$$\bar{F}^*_{Z_n}(z) = F^*_{W_n}(e^{-\lambda z}). \tag{21}$$

Per tant, de manera natural, l'estimador de $\bar{F}_X(x)$ és

$$\bar{F}^*_{X_n}(x) = \bar{F}^*_{Z_n}(x) - \frac{f^*_Z(x)}{\lambda}, \tag{22}$$

on

$$f^*_Z(x) = \lim_{h \to 0^+} \frac{\bar{F}^*_{Z_n}(x) - \bar{F}^*_{Z_n}(x-h)}{h}.$$

En aquest cas $\bar{F}_{X_n}(x)$ és el NPMLE de $\bar{F}_X(x)$ ja que $\bar{F}^*_{Z_n}$ i f^*_Z són els NPMLE de $\bar{F}_Z$ i f_Z, respectivament.

En el cas que les dades completes $\{W_i\}_{i=1}^n$ no siguin disponibles, sinó que les dades tinguin casos censurats o perduts, els autors recomanen utilitzar l'estimador de Kaplan-Meier en comptes de la funció de distribució empírica per a obtenir $F^*_{W_n}$.

3. Mètode de Xu et al.

3.1. Resum. A Xu et al. [**13**], els autors exploren la metodologia presentada per Xu & Prorok [**11**] formalitzant la hipòtesi de dependència entre el *lead-time* i el *post-lead-time*, basant-se en l'assumpció que és biològicament raonable una correlació positiva entre aquests dos temps. Expressen el *post-lead-time* a partir del *lead-time* i uns paràmetres auxiliars per tal d'assegurar la correlació positiva entre els temps. Dels resultats obtinguts, es pot inferir que l'assumpció d'independència condueix a una sobreestimació de la supervivència del *post-lead-time*.

3.2. Model. Com en Xu & Prorok [**11**], el model assumeix que la història natural de la malaltia correspon a la d'una malaltia progressiva amb transicions $S_0 \to S_p \to S_c$. Denoten Z com el temps total de supervivència des del diagnòstic per cribratge com la suma del *lead-time* Y i el temps de supervivència extra o *post-lead-time* X.

$$Z \stackrel{d}{=} X + Y, \tag{23}$$

Tant Y com X són inobservables ja que el punt del diagnòstic clínic ($S_p \to S_c$) d'un individu detectat per cribratge és generalment desconegut. Per això, les observacions disponibles per a un individu del grup detectat per cribratge és Z. Com citen Xu & Prorok [**11**], el model 23 és un model de deconvolució assumint que Y és totalment coneguda. En aquest cas fan una assumpció paramètrica sobre Y, assumint que és distribueix exponencialment amb paràmetre λ conegut. El model també ignora el fet que el tractament posterior a la detecció per cribratge comporti un risc de mort més gran en comparació amb el fet de no haver estat diagnosticat.

L'assumpció de dependència entre el *lead-time* i el *post-lead-time* és crucial, per tal de modelitzar aquesta relació expressen el *post-lead-time* com

$$\begin{aligned} &X = (1-c)Y + \delta\tau, \quad \tau, \delta \text{ i } Y \text{ independents} \\ &P(\tau \geq 0) = 1 \\ &P(\delta = 1) = 1 - P(\delta = 0) = c \,(\text{conegut}). \end{aligned} \tag{24}$$

Segons els autors, la motivació d'aquesta relació està basada en la presumpció que és biològicament raonable pensar que el *lead-time* i el *post-lead-time* són proporcionals i positivament correlacionats. Amb aquest model obtenen una manera senzilla d'aconseguir aquesta relació, permetent una estimació no paràmetrica.

La correlació entre X i Y es calcula amb

$$E(X) = (1-c)E(Y) + cE(\tau)$$

$$E(XY) = (1-c)E(Y^2) + cE(\tau)E(Y)$$

el que implica que

$$\begin{aligned} Cov(X,Y) &= (1-c)Var(Y) \\ \rho_c = Corr(X,Y) &= \tfrac{1-c}{\lambda\sqrt{Var(X)}} \end{aligned} \tag{25}$$

sempre que $E(T^2) < \infty$. Quan $P(\delta = 1) = 1$, és té el cas particular del model presentat per Xu & Prorok [**11**] on s'assumeix independència entre el *lead-time* i el *post-lead-time*. Generalitzant aquest model amb les noves variables τ i δ, s'obté

$$\frac{Z}{2-c} \overset{d}{=} Y + \frac{\delta\tau}{2-c} \tag{26}$$

Aplicant els estimadors màxim versemblants obtinguts en Xu & Prorok [**11**] considerant $Z/(2-c)$ i $\delta\tau/(2-c)$ com el temps total de supervivència i el *post-lead-time* del model 23, respectivament, s'arriba a

(27)

$$F^*_{Z_n}(z) = \bar{F}^*_{W_n}(e^{-\lambda z/(2-c)})$$

$$f^*_Z(z) = \lim_{h\to 0^+} \frac{\bar{F}^*_{Z_n}(z) - \bar{F}^*_{Z_n}(z-h)}{h}$$

$$\bar{F}^*_{X_n}(x) = e^{-\lambda x/(1-c)} \left\{ 1 + \frac{\lambda}{1-c} \int_0^x \left[\bar{F}^*_{Z_n}(z) - \frac{2-c}{\lambda} f^*_Z(z) \right] e^{-\lambda z/(1-c)}\, dz \right\}$$

on $\bar{F}^*_{Z_n}(z) = 1 - F^*_{Z_n}(z)$, $\bar{F}^*_{W_n}(w) = 1 - F^*_{W_n}(w)$ i

$F^*_{W_n}(w) =$ menor majorant còncava de la funció de distribució empírica $F_{W_n}(w)$ de $W_1, ..., W_n$,

$$f^*_{W_n}(w) = \lim_{h\to 0^+} \frac{F^*_{W_n}(w) - F^*_{W_n}(w-h)}{h},\ w < W_{(n)}$$

$$W_i = \exp\left[-\lambda Z_i/(2-c)\right],\ i = 1, ..., n$$

$$W_{(n)} = \max(W_i, ..., W_n).$$

Per tant, $f^*_{W_n}(w)$ és el pendent de la menor majorant còncava de $F^*_{W_n}(w)$ a w [**12**]. Es pot obtenir $F^*_{W_n}(w)$ utilitzant línies rectes per a connectar punts de salt apropiats de la funció de distribució empírica $F_{W_n}(w)$ de tal manera que $F^*_{W_n}(w)$ és la funció de distribució menor concàva respecte a $F_{W_n}(w)$.

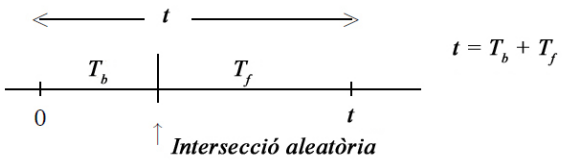
En el cas que les dades completes $\{W_i\}_{i=1}^n$ no siguin disponibles, sinó que hi hagi casos censurats, els autors recomanen utilitzar l'estimador de Kaplan-Meier en comptes de la funció de distribució empírica per a obtenir $F^*_{W_n}$.

4. Mètode de Zelen

4.1. Resum. Marvin Zelen [**9**], amplia els models de detecció precoç (secció 4 del capítol 2) incorporant models específics per edats. Defineix el *backward recurrence time* com la quantitat de temps que un individu ha tingut la malaltia pre-clínica abans de ser detectada per examen, i el *forward recurrence time* com el *lead-time*. En aquest model la probabilitat de transició a l'estat pre-clínic pot estar relacionada amb l'edat, resultant en un procés no estacionari. Els resultats obtinguts són les distribucions per als diferents *recurrence time* i la seva generalització per a mostres afectades pel biaix de durada (*length biased sampling*).

4.2. Model. El model presentat assumeix una història natural de la malaltia $S_0 \to S_p \to S_c$. Cal notar que la transició $S_0 \to S_p$ mai és observada, i la transició $S_p \to S_c$ descriu la incidència de la malaltia. El principal objectiu dels programes de detecció precoç és diagnosticar els individus en l'estat pre-clínic mitjançant un examen especial. Un cop diagnosticada en l'estat pre-clínic, la malaltia serà tractada i l'història natural de la malaltia es veurà alterada. Com a resultat, la transició $S_p \to S_c$ mai serà observada. Zelen defineix el temps guanyat pel diagnòstic precoç com el *forward recurrence time* i el temps que una persona ha estat en l'estat pre-clínic abans del diagnostic precoç com el *backward recurrence time* (Figura 3.3).

FIGURA 3.3. Forward i backward recurrence time



L'autor defineix la probabilitat de transició $S_0 \to S_p$ durant $(\tau, \tau + d\tau)$ com $w(\tau)d\tau$, on τ és l'edat d'entrada a S_p.

La incidència puntual es refereix a la transició $S_p \to S_c$, i la relació entre $w(z)$ i $I(z)$ és

$$I(z) = \int_0^z w(\tau) q(z-\tau)\, d\tau, \tag{28}$$

on $q(t)$ és la funció de densitat de probabilitat del temps de sojorn en S_p.

Assumint que la malaltia és detectada precoçment en el temps cronològic t_0 i que l'edat en el moment de la detecció és z, l'autor obté l'expressió de la distribució del *forward recurrence time* a partir de

$$P(z)Q_f(t|z) = \int_0^z w(\tau)Q(z-\tau+t)\,d\tau, \tag{29}$$

on $P(z) = \int_0^z w(\tau)Q(z-\tau)\,d\tau$.

Una expressió que li permetrà calcular la funció de densitat

$$q_f(t|z) = -\frac{d}{dt}Q_f(t|z) = \int_0^z w(\tau)q(z-\tau+t)\,d\tau/P(z). \tag{30}$$

Llavors aplicant el mateix raonament per al *backward recurrence time*, obté l'expressió de la distribució

$$P(z)Q_b(t|z) = \int_0^{z-t} w(\tau)Q(z-\tau)\,d\tau,\ 0 < t \leq z \tag{31}$$

que permet el càlcul de la funció de densitat del *backward recurrence time*

$$q_b(t|z) = -\frac{d}{dt}Q_b(t|z) = w(z-t)Q(t)/P(z),\ 0 < t \leq z \tag{32}$$

De forma similar l'autor obté la distribució del temps de sojorn en S_p, definida com $T = T_b + T_f$. Com que l'edat z està fixada en el moment de la detecció t_0, considera els casos $t > z$ i $t \leq z$ respectivament. Per al cas de t fixada i $t < z$ és necessari que $z - t < \tau < z$, on τ és l'edat d'entrada a pre-clínic. Utilitzant el resultats obtinguts en el capítol 2, Zelen deriva la funció de densitat del temps de sojorn

$$f(t|z) = \begin{cases} q(t)\int_0^z w(\tau)d\tau/P(z) & t > z \\ q(t)\int_{z-t}^z w(\tau)d\tau/P(z) & t \leq z \end{cases} \tag{33}$$

5. Altres mètodes

5.1. Mètode de Mahnken et al. El mètode presentat per Mahnken et al. [**14**] proposa un model conceptual dividint la duració de la malaltia en dues fases: preclínica (T_0) i simptomàtica (T_1). El resultat d'interès és el vector bivariant (T_0, T_1), aquest model permet tenir en compte els biaixos presents en la detecció precoç. A partir de les dades observades es defineixen diferents regions de censura per a T_0 i T_1

i es construeix la funció de versemblança, d'on s'obtenen les estimacions necessàries per a inferir sobre l'efecte dels programes de detecció precoç.

5.2. Mètode de Duffy et al. En aquest mètode[**15**], els autors apliquen una correcció simple al *lead-time* esperat, condicionat-lo a què sigui més petit que el temps de supervivència i assumint una distribució exponencial del temps de sojorn. Pel que fa al *length bias* realitzen un anàlisi de sensibilitat per estimar les probabilitats de ser detectat mitjançant el cribatge i de morir de càncer de mama durant el període d'observació i el risc relatiu de morir de càncer de mama entre les dones detectades per cribatge i no cribratge independentment del *length bias*.

Capítol 4
Mètodes

1. Dades de supervivència

S'han utilitzat dades de supervivència de càncer de mama provinents dels registres de càncer poblacionals de les províncies de Girona i Tarragona. Com aquests dos registres cobreixen un 20% de la població catalana, s'han combinat per a obtenir dades de supervivència per a Catalunya. Per altra banda, també s'han inclòs les dades de supervivència del programa de detecció precoç de l'Hospital del Mar. En aquest secció, es presenta l'anàlisi descriptiva de les dades i l'estudi de la supervivència específica i relativa del càncer de mama.

1.1. Dades dels registres.

1.1.1. *Dades dels registres poblacionals de càncer de Catalunya.* A Catalunya existeixen dos registres de càncer de base poblacional, el de Girona i el de Tarragona.

El Registre de Càncer de Girona (RCG) és un registre de base poblacional que recull informació sobre tots els casos de càncer de la Regió Sanitària de Girona des de l'any 1994. Cobreix una població aproximada de 707.000 habitants (estimacions postcensals a 31 de desembre de l'any 2008 de l'Institut d'Estadística de Catalunya, IDESCAT), residents a les comarques del Gironès, La Selva, Alt Empordà, Baix Empordà, Garrotxa, Ripollès i Pla de l'Estany. El RCG fou establert l'any 1994 com ampliació del Registre monogràfic de càncer de mama i genital femení que va estar en funcionament des de l'any 1980 fins l'any 1989. L'encarregat de gestionar-lo és la Unitat d'Epidemiologia i Registre del Càncer del Pla Director d'Oncologia.

El Registre de Càncer de Tarragona es va fundar l'any 1979 i disposa de dades poblacionals des de 1980. Cobreix una població aproximada de 785.133 habitants (estimacions postcensals a 31 de desembre de l'any 2008 de l'IDESCAT), residents a les comarques de l'Alt Camp, Baix Camp, Baix Ebre, Baix Penedès, Conca de Barberà, Montsià, Priorat, Ribera d'Ebre, Tarragonès i Terra Alta. L'encarregat de gestionar-lo és la Fundació Lliga per a la Investigació i Prevenció del Càncer

(FUNCA), amb la col·laboració del Departament de Salut i el Servei Català de la Salut.

Les dades dels registres de càncer de Girona i Tarragona, amb les dades de mortalitat per càncer del registre de mortalitat de Catalunya i les piràmides poblacionals de Catalunya proporcionades per l'IDESCAT han permès obtenir les estimacions dels indicadors de càncer per a Catalunya (incidència, mortalitat i supervivència).

S'han combinat les dades dels casos de càncer de mama de les províncies de Girona, diagnosticats entre els anys 2002 i 2006 i de Tarragona diagnosticats entre 2000 i 2005. S'han obtingut un total de 3972 dones diagnosticades de càncer de mama, 1783 a Girona i 2189 a Tarragona.

Tot i que tenim períodes de seguiment diferents, suposarem que no hi ha diferències entre aquests períodes i reescalarem el temps d'inici individuals a 0. En el cas de Girona el temps de seguiment és fins al 31 de desembre de 2009 i pel cas de Tarragona és fins al 31 de desembre de 2008.

1.1.2. *Dades del Programa de detecció precoç de càncer de mama de l'Hospital del Mar.* El programa de detecció precoç de l'Hospital del Mar, que és el més antic existent a Catalunya i que ja registra més de 15 anys d'història, ha permès fer un estudi evolutiu del càncer de mama i obtenir dades concretes sobre tots els tipus i subtipus de tumors detectats gràcies al cribratge poblacional.

Des què el 1995 l'Hospital del Mar va posar en marxa un programa de detecció precoç de càncer de mama, s'han fet més de 210.000 mamografies que han permès la detecció precoç de més de 900 casos de càncer de mama. En un inici es va convidar a les dones de 50 a 64 anys dels districtes municipals de Ciutat Vella i Sant Martí a fer-se una mamografia cada dos anys. L'Hospital de l'Esperança l'any 1999 va iniciar el Programa de Detecció Precoç al districte municipal de Gràcia i l'any 2001 al districte municipal de Sarrià-Sant Gervasi. A partir de l'any 2004 tota la població de Barcelona té la possibilitat de participar en aquest programa de detecció precoç en diferents centres sanitaris.

El nombre de casos de dones diagnosticades de càncer de mama entre els anys 1996 i 2005 va ser de 1704 dones. En aquest cas la data de la fi del seguiment és el 31 de desembre de 2007.

1.2. Anàlisi descriptiva. Per a la nostra anàlisi s'ha considerat un conjunt de variables relacionades amb la detecció precoç i la mortalitat. Per als registres de càncer de Girona i Tarragona s'han estudiat les següents variables:

- **Edat al diagnòstic**: Variable contínua categoritzada per a propòsits descriptius en menors de 40 anys, entre 40 i 50 anys, entre 50 i 60 anys, entre 60 i 70 anys i més grans de 70 anys.

- **Tipus de diagnòstic**: Diagnosticades per cribratge vs no diagnosticades per cribratge.
- **Estat al final del seguiment**: Viva vs morta.
- **Mort càncer de mama**: Mort càncer de mama vs mort altres causes.

En l'anàlisi de les dades de l'Hospital del Mar s'han considerat les següents variables:

- **Edat al diagnòstic**: Variable contínua categoritzada per a propòsits descriptius en menors de 40 anys, entre 40 i 50 anys, entre 50 i 60 anys, entre 60 i 70 anys i més grans de 70 anys.
- **Tipus de diagnòstic**: Diagnosticades per cribratge vs no diagnosticades per cribratge.
- **Estadi**: Grau d'extensió del tumor en el moment del diagnòstic. Considerant Estadi I, Estadi II, Estadi III i Estadi IV.
- **Estat al final del seguiment**: Viva vs morta.
- **Mort càncer de mama**: Mort càncer de mama vs mort altres causes.

S'ha analitzat si hi ha diferències entre les dones diagnosticades precoçment i les detectades simptomàticament utilitzant la prova exacta de Fisher.

1.3. Anàlisi de la Supervivència. Per a determinar la proporció de dones que sobreviuen al càncer de mama, suposant que aquesta sigui l'única causa de mort, hem estimat la supervivència per causa específica. Amb aquest mètode hem considerat la mort per càncer de mama com l'esdeveniment d'interès. Les morts per altres causes o pèrdues s'han tractat com a observacions censurades per la dreta. S'ha assumit que la censura és no informativa. El temps de supervivència s'ha calculat fent la diferència entre la data del diagnòstic i la data de defunció per càncer de mama, altrament seria la data de tancament del seguiment (31/12/2008 o 31/12/2009 segons el registre poblacional i 31/12/2007 pel registre del programa) i estaria censurat per la dreta.

Taula 4.1. Censura i temps de seguiment

Variable	**% Censura**	**Mediana**
Girona i Tarragona	85.3%	4.88 anys
Hospital del Mar	93.4%	4.86 anys

La Taula 4.1 mostra el percentatge de censura i la mediana del temps de seguiment. Les nostres dades presenten un gran percentatge de casos censurats i això dificultarà la interpretació dels resultats.

El mètode que s' utilitza no té en compte altres causes de mortalitat en competició i pot originar una subestimació de la supervivència. Habitualment, el què fan els

registres de càncer és calcular la supervivència relativa, és a dir, estimar la supervivència global del càncer de mama (totes les morts es tracten igual) i corregir-ho per la supervivència de la població general de la zona d'estudi. La supervivència relativa s'obté dividint la supervivència observada entre l'esperada segons la mortalitat de la població resident en l'àrea geogràfica de l'estudi. Els resultats de supervivència relativa s'han obtingut amb l'aplicació web per al càlcul de la supervivència relativa WAERS de l'Institut Català d'Oncologia (ICO) [**16**], prenent com a població de refèrencia tota Catalunya en cas dels registres de càncer i la població de la província Barcelona per al programa de cribratge.

2. Estudi de Simulació

S'ha elaborat un estudi de simulació del seguiment d'una cohort de dones a partir de dades reals d'incidència, mortalitat i supervivència. Amb l'ajuda del paquet estadístic R, s'han simulat casos de dones diagnosticades simptomàticament de càncer de mama, i per a tots aquests casos s'han generat temps de supervivència i temps de sojorn en estat pre-clínic, és a dir, la seva història natural de la malaltia.

Hem obtingut estimacions del *lead-time bias* i el *length bias* a partir dels temps generats i la simulació de diferents patrons de programes de detecció precoç (edats d'inici i final i periodicitat de les proves).

2.1. Disseny de la simulació. Els principals motius per a dur a terme aquest estudi de simulació han estat, d'una banda, intentar conèixer a fons el fenòmen de la detecció precoç del càncer de mama i, de l'altra, els pocs casos de dones mortes per càncer de mama que trobem a les dades dels registres. S'ha dissenyat un estudi de simulació el més realista possible. Per això s'han utilitzat les dades d'incidència, mortalitat i supervivència de Catalunya obtingudes en diferents estudis del grup de recerca [**17, 18, 19**].

En general, no es disposa d'informació sobre la fase pre-clínica de la malaltia (S_p). Només a partir de diferents assaigs clínics s'han pogut estimar paràmetres com les probabilitats de transició a S_p i la distribució del temps de sojorn en estat pre-clínic. Per aquest fet, s'ha considerat convenient simular en un primer moment la fase clínica de la malaltia (S_c) a partir d'estimacions de la incidència, la supervivència i la mortalitat de càncer de mama i altres causes de mort de Catalunya. Amb aquestes funcions s'ha obtingut l'edat d'inici de la fase clínica, la durada dels temps de supervivència i la causa de mort de les dones diagnosticades de càncer de mama.

Un cop simulada la fase clínica, per a cada dona incident de càncer de mama s'ha simulat la fase pre-clínica a partir dels paràmetres obtinguts de la bibliografia. Per tal de descriure la dependència del temps de supervivència i el temps de sojorn s'ha utilitzat un model de còpula de Clayton. Per tal de modelitzar matemàticament la simulació, s'han definit les variable aleatòries $X(>0)$ com l'edat de la dona i

$T(>0)$ com el temps de supervivència des del diagnòstic fins a la mort de la dona per càncer de mama o altres causes.

2.1.1. *Escenaris de simulació.* S'han considerat diferents patrons de programa de detecció precoç en funció de l'edat d'inici i final i la periodicitat dels exàmens. Participen en el programa totes les dones vives que no han estat diagnosticades de càncer de mama abans de l'edat d'inici del programa. S'han generat $B = 100$ conjunts de dades amb temps clínic i pre-clínic de dones incidents de càncer de mama. Per a cada conjunt de dades, s'han aplicat diferents patrons de cribratge. Els escenaris estan determinats pels següents paràmetres

- **Mida de la cohort**, N
  Nombre de dones que conformen la cohort d'estudi, $N = 100000$
- **Cohort específica**, ν
  Any de naixement de la cohort. S'ha considerat la cohort $\nu = 1950$ com la més representativa i la més treballada pel nostre grup de recerca.
- **Edat d'inici del cribratge**, *ecrib*
  Edat, en anys, d'inici del programa de detecció precoç, $ecrib \in \{40, 50\}$.
- **Periodicitat entre exàmens**, *tex*
  Temps, en anys, entre cada mamografia que es realitzarà durant el programa de detecció precoç, $tex \in \{1, 2\}$

A la Taula 4.2 es presenten els diferents escenaris:

TAULA 4.2. Escenaris de la simulació

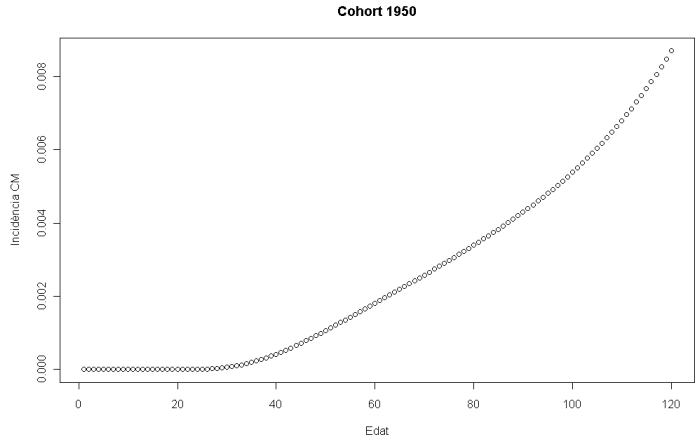
Escenaris	N	ν	**Edats exàmens**	**Nº exàmens**	**Periodicitat**
Escenari 1	100000	1950	40-68	15	2 anys
Escenari 2	100000	1950	40-69	30	1 anys
Escenari 3	100000	1950	50-68	10	2 anys
Escenari 4	100000	1950	50-69	20	1 anys

2.1.2. *Dades d'incidència, mortalitat i supervivència.* Les dades d'incidència de càncer de mama s'han obtingut a partir de models d'incidència de càncer de mama estimats en treballs anteriors amb dades de Catalunya [**17**]. S'ha utilitzat el model d'incidència de càncer de mama en absència de cribratge (*background*), al qual posteriorment s'han aplicat els diferents escenaris de cribratge. Aquest model d'incidència depèn de l'edat al diagnòstic i de la cohort de naixement.

Normalment les dades d'incidència estan agrupades per edats, on el j-èssim interval d'edat és $A_j = (a_{j-1}, a_j]$ per $j = 1, .., 120$, amb $a_0 = 0$. S'han considerat intervals constants amb $\Delta = a_j - a_{j-1} = 1$. Sigui $I(x)$ la probabilitat de ser incident durant l'interval d'edat $(x, x + dx)$, la incidència per al j-èssim grup d'edat s'obté per $I_j = \int_{A_j} I(x)dx$. A la Figura 4.1 es presenten les taxes d'incidència de càncer de mama per 100.000 dones, agrupades per intervals d'edat A_j per a la cohort

$\nu = 1950$. S'han aplicat aquestes taxes d'incidència a la cohort i s'ha obtingut els nombre de dones incidents de càncer de mama per edats.

FIGURA 4.1. Taxes d'incidència de càncer de mama específiques per intervals d'edat



S'han tingut en compte les altres causes de mortalitat en competició durant el seguiment de la cohort. Per aquest fet, per a la població general s'ha obtingut la taxa de mortalitat per altres causes a partir de les taxes de mortalitat global i la taxa de mortalitat de càncer de mama a Catalunya [**18**]. Per estimar la funció de risc específica de mort per altres causes, $\lambda_{AC}(x)$, s'ha utilitzat la taxa anual de mortalitat per altres causes, per grups d'edat de la cohort. A la Figura 4.2 tenim representades les taxes de mortalitat per altres causes, M_j, per 100.000 dones, agrupades a partir dels intervals d'edat A_j per a una cohort $\nu = 1950$.

Per als temps de supervivència des del diagnòstic, s'han utilitzat les funcions de supervivència de càncer de mama de les dones no diagnosticades abans que s'utilitzés la mamografia de cribratge, estimades a partir de les dades observades entre 1980 i 1989 a Catalunya i entre 1975 i 1979 als Estats Units [**19**]. S'han utilitzat les funcions de densitat de probabilitat, $f_{CM}(t)$, per grups d'edat d'incidència, ponderades per la distribució d'estadis del càncer de mama al moment del diagnòstic. Els grups d'edat estudiats han estat 25-39 anys, 40-49 anys, 50-59 anys, 60-69 anys i 70-85 anys. Per a les dones que s'han posat incidents en edats abans dels 25 anys o després dels 85, s'han utilitzat les funcions de supervivència del grup d'edat més proper. A la Figura 4.3 tenim representades les funcions de densitat pels diferents grups d'edat corresponents a 25 anys de seguiment.

2.1.3. *Sensibilitat de la mamografia i temps de sojorn a S_p.* La sensibilitat dels exàmens de detecció precoç i la distribució del temps de sojorn han estat extretes de les publicacions de Lee i Zelen [**20**]. Els autors assumeixen que el temps de sojorn

FIGURA 4.2. Funció de risc específica de mort per altres causes agrupada per intervals d'edat

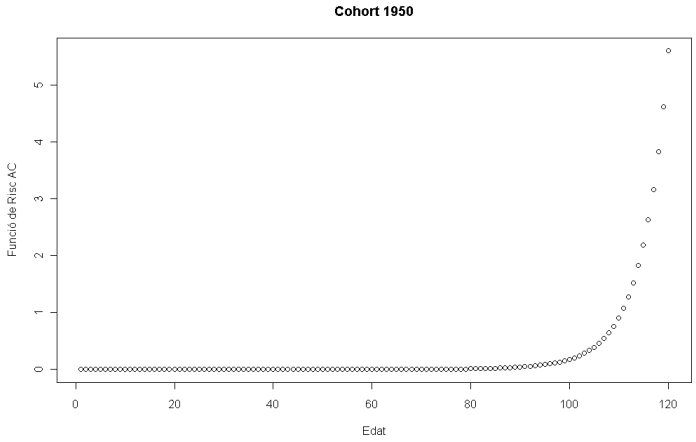
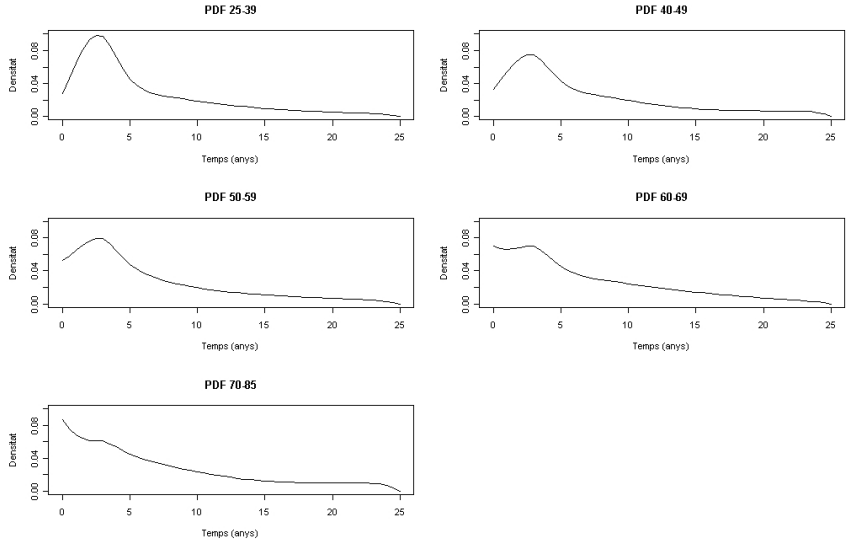


FIGURA 4.3. Funcions de densitat de supervivència de càncer de mama ponderades per estadis per als diferents grups d'edat



en estat pre-clínic segueix una distribució exponencial amb una mitjana dependent de l'edat d'entrada a l'estat pre-clínic. Llavors, els temps de sojorn esperats, $m(x)$, són:

$$m(x) = \begin{cases} 2 & x \leq 40 \\ -6 + 0.2x & 40 < x \leq 50 \\ 4 & x > 50 \end{cases} \tag{34}$$

Aquests valors estan basats en dades de diferents assaigs clínics aleatoritzats sobre detecció precoç. Pel que fa a les dades de sensibilitat de la mamografia, Lee i Zelen van obtenir-les d'estimacions publicades pel Breast Cancer Surveillance Consortium (BCSC), un projecte per tal d'avaluar la pràctica de la mamografia de cribratge entre la població d'Estats Units. Aquestes sensibilitats, $\beta(x)$, són dependents de l'edat de la dona a l'examen de cribratge i són les següents:

$$\beta(x) = \begin{cases} 0.55 & x < 40 \\ 0.65 & 40 \leq x < 45 \\ 0.70 & 45 \leq x < 50 \\ 0.75 & 50 \leq x < 70 \\ 0.80 & x \geq 70 \end{cases} \tag{35}$$

2.1.4. *Càlcul de la funció de supervivència global.* Amb les funcions de densitat presentades a la secció 2.1.2 es podrien obtenir els temps de supervivència específica de càncer de mama de cada grup d'edat a partir de la relació

$$S_{CM}(t) = \int_t^\infty f_{CM}(u)\, du. \tag{36}$$

Aplicant el mètode de la transformada inversa s'obtindrien els temps $T = S_{CM}^{-1}(u)$, on u és una variable aleatòria uniforme en l'interval $(0, 1)$, $U(0, 1)$.

Tot i així, haurem de tenir en compte que la mortalitat per altres causes està actuant al mateix temps que la mortalitat per càncer de mama, en un context de riscos competititius. Per tant, és més adient estimar la funció de supervivència global, $S(t)$, a partir de les supervivències específiques.

En primer lloc, s'ha calculat la funció de supervivènca específica per altres causes a partir de $\lambda_{AC}(x)$. Atès que el suport de la funció de risc correspon a l'edat de la dona, s'ha d'utilitzat una distribució del temps de supervivència de mort per altres causes condicionada a l'edat d'incidència.

La distribució condicionada d'una variable aleatòria T donada una edat $X = x$ és coneix com la distribució del *future lifetime* [**21**]. La seva funció de densitat de probabilitat és

$$f_T(t) = \frac{f_X(x+t)}{S_X(x)}, \quad t > 0, \tag{37}$$

i la seva funció de supervivència és

$$S_T(t) = \frac{S_X(x+t)}{S_X(x)}, \quad t > 0, \tag{38}$$

llavors la funció de risc corresponent és

$$\lambda_T(t) = \lambda_X(x+t), \quad t > 0, \tag{39}$$

Per al cas de la supervivència específica de morir per altres causes, s'ha calculat la funció de supervivència, $S_{AC}(x)$, que correspon a l'edat d'incidència a partir de la relació

$$S_{AC}(x) = \exp\{-\Lambda_{AC}(x)\} = \exp\{-\int_0^x \lambda_{AC}(u)\, du\} \tag{40}$$

on $\Lambda_{AC}(x)$ és la funció de risc acumulada [**22**]. Llavors, s'obté la funció de supervivència del *future lifetime* per altres causes amb l'expressió

$$S_{AC}(t) = \frac{S_{AC}(x+t)}{S_{AC}(x)}, \quad t > 0, \tag{41}$$

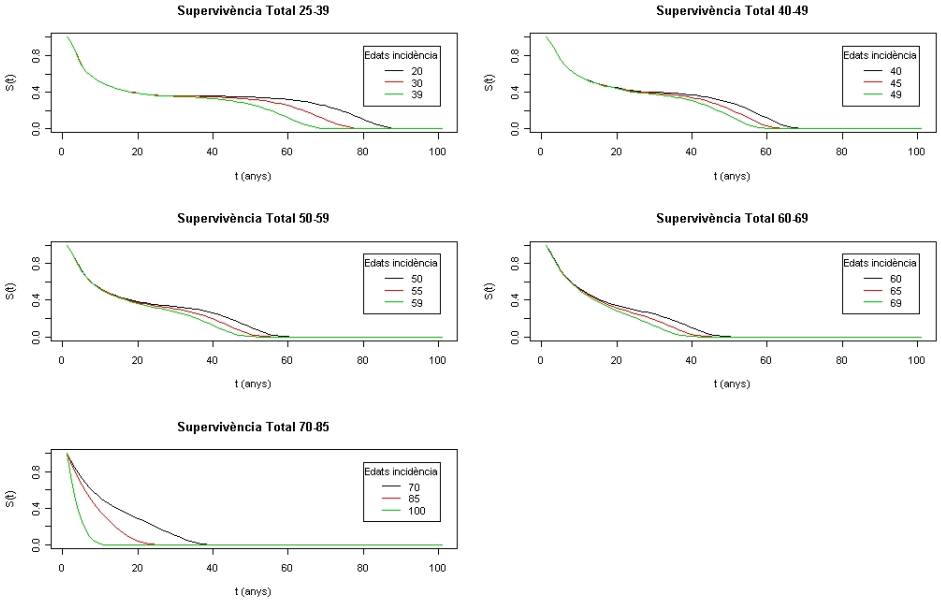
La funció de supervivència global es pot expressar com el producte de les funcions de supervivència específiques.

$$S(t) = S_{AC}(t) S_{CM}(t), \quad t > 0, \tag{42}$$

Per a cada grup d'edat d'incidència del càncer de mama s'ha estimat la funció de supervivència associada. Tal com s'ha construït la funció, aquesta varia per a cada dona incident del grup d'edat, ja que la funció de supervivència del *future lifetime* per altres causes depèn de l'edat d'incidència. En les següents seccions s'explica com s'ha realitzat la simulació dels temps de supervivència i com s'ha assignat la causa de mort. A la Figura 4.4, es presenten les funcions de supervivència global per als diferents grups amb diferents edats d'incidència per a la cohort $\nu = 1950$. El temps de supervivència s'ha definit com el temps entre moment del diagnòstic i la mort per càncer de mama o per altres causes.

Es pot observar com per als grups d'edats més joves, les morts de càncer de mama fan baixar la supervivència ràpidament en els primers anys després del diagnòstic fins que s'estabilitza a partir dels 25 anys. Més endavant, les morts per altres causes fan caure la supervivència a zero. Es pot veure com a mesura que augmenta l'edat d'incidència, els temps de supervivència son més curts ja que les dones diagnosticades en edats avançades tenen més risc de morir per altres causes.

FIGURA 4.4. Funcions de supervivència global de la població incident per als diferents grups d'edat



2.2. Models còpula per a la simulació de dades de supervivència bivariant. Hi ha evidències biològiques que mostren una correlació positiva de la fase clínica amb la fase pre-clínica. Un temps de sojorn curt en estat pre-clínic implica que la malaltia és agressiva mentre que un temps de sojorn llarg implica que la malaltia té un creixement lent. Per tant, aquelles dones amb llargs temps de sojorn en estat pre-clínic tendiran a tenir una durada clínica llarga i viceversa.

Per tal de descriure la relació de depèndencia entre el temps de sojorn en estat preclínic i el temps de supervivència en estat clínic s'han utilitzat els models còpula per a dades de supervivència bivariants. Les còpules permeten modelitzar de forma separada les distribucions marginals i l'estructura d'associació. Els models còpula assumeixen que les distribucions marginals no depenen de l'estructura de dependència [**23, 24**]. A continuació es defineixen les funcions còpules i les seves propietats.

2.2.1. *Funció Còpula.* Una còpula $C_\alpha(u, v)$ és una funció contínua bivariant definida com

$$C_\alpha(u, v) : [0, 1] \times [0, 1] \to [0, 1], \tag{43}$$

de manera que és no decreixent per a cada component de (u, v) i satisfà $C_\alpha(u, 0) = C_\alpha(0, v) = 0$ i $C_\alpha(u, 1) = u$, $C_\alpha(1, v) = v$. La forma de la funció $C_\alpha(u, v)$ depèn d'un vector de paràmetres $\alpha' = (\alpha_1, ..., \alpha_p)$.

Siguin T_1 i T_2 dues variables aleatòries no negatives amb funcions de supervivència marginals $S_1(s)$ i $S_2(t)$. Sigui $C_\alpha(u, v)$ una funció còpula per $0 \leq u \leq 1$ i $0 \leq v \leq 1$ i on α mesura l'associació entre T_1 i T_2. S'assumeix que la funció de supervivència conjunta de (T_1, T_2), $S(s, t)$, pot ser expressada com a funció de les marginals S_1 i S_2 i el paràmetre α mitjançant la següent expressió

$$S(s, t) = C_\alpha(S_1(s), S_2(t)). \tag{44}$$

Es compleix que $S(s, \infty) = 0$, $S(\infty, t) = 0$ i

$$\begin{aligned} S(s, 0) &= C_\alpha(S_1(s), 1) = S_1(s) \\ S(0, t) &= C_\alpha(1, S_2(t)) = S_2(t). \end{aligned} \tag{45}$$

2.2.2. *Mesures de dependència.* La dependència entre T_1 i T_2 pot ser descrita mitjançant mesures de concordància com el coeficient de correlació d'Spearman, ρ_S, o el coeficient de concordància de Kendall, τ_K. El coeficient de correlació d'Spearman és una mesura de la correlació de rangs entre dues variables aleatòries contínues, mentre τ_K mesura la concordància entre dos parells (T_{1i}, T_{2i}) i (T_{1j}, T_{2j}).

L'ús del coeficient de correlació lineal de Pearson, ρ, directament a (T_1, T_2) és inapropiat des del moment que aquesta mesura es limita a variables aleatòries amb una distribució normal bivariant, poc habitual en el context de variables aleatòries de temps de vida.

Tots dos coeficients, ρ_S i τ_K poden ser expressats en termes d'una funció còpula:

$$\begin{aligned} \rho_S &= 12 \int_0^1 \int_0^1 C_\alpha(u, v) du dv - 3 \\ \tau_K &= 4 \int_0^1 \int_0^1 C_\alpha(u, v) dC_\alpha(u, v) - 1 \end{aligned} \tag{46}$$

Per simplicitat, s'ha utilitzat principalment el coeficient de concordància de Kendall τ_K.

2.2.3. *Còpula de Clayton.* Sigui ϕ_α una funció convexa decreixent definida en $(0, 1]$ tal que $\phi_\alpha(1) = 0$. Una funció còpula arquimediana de $(u, v) \in [0, 1]^2$ ve determinada per

$$C_\alpha(u,v) = \phi_\alpha^{-1}\left\{\phi_\alpha(u) + \phi_\alpha(v)\right\}, \tag{47}$$

Les còpules arquimedianes són una important família de còpules, que tenen una forma simple, amb característiques com ara l'associativitat i una gran varietat d'estructures de dependència.

La còpula de Clayton és un cas especial de còpula arquimediana amb $\phi_\alpha(x) = (x^{1-\alpha} - 1)/(\alpha - 1)$. La funció còpula ve explícitament determinada per

$$C_\alpha(u,v) = \left\{u^{1-\alpha} + v^{1-\alpha} - 1\right\}^{1/(1-\alpha)}, \tag{48}$$

amb $\alpha > 1$, $(u,v) \in [0,1]^2$. Sovint s'utilitza una altra parametrització amb $\theta = \alpha - 1$.

Un dels motius pels quals s'ha escollit la còpula de Clayton per a descriure la dependència del temps de sojorn i el temps de supervivència és que permet associacions positives entre els temps. Quan $\alpha \to 1^+$, $S(s,t) \to S_1(s)S_2(t)$, hi ha independència entre T_1 i T_2, i quan $\alpha \to \infty$, $S(s,t) \to \min\{S_1(s), S_2(t)\}$, la distribució bivariant que mostra associació màxima entre T_1 i T_2.

Sota aquest model, la tau de Kendall és igual a $\tau_K = \frac{\alpha-1}{\alpha+1}$. Amb aquesta propietat de la còpula de Clayton, es poden reproduïr diferents estructures de dependència a partir de valors d'α.

2.3. Generació dels conjunts de dades. Per a cada conjunt de dades d'un escenari particular s'han generat n dones incidents de càncer de mama i s'han recopilat les seguents variables:

- X_S: edat d'entrada a S_p
- T_S: durada del temps de sojorn en S_p
- X_C: edat d'entrada a S_c
- T_C: durada del temps de supervivència en S_c
- D_{CM}: edat de mort per càncer de mama
- D_{AC}: edat de mort per altres causes
- X_L: edat de la detecció precoç en S_p
- T_L: durada del *lead-time*
- C: causa de la mort (1 = Càncer de Mama, 0 = Altres causes).

2.3.1. *Generació d'n i d'X_C.* S'ha considerat que la incidència i la mortalitat per altres causes actuen al mateix temps durant el seguiment de la cohort. S'han generat casos de dones incidents i mortes fins que no ha quedat cap dona susceptible de morir per altres causes o ser diagnosticada de càncer de mama.

S'ha definit la variable nombre de dones vives, N_j, per al j-èssim interval d'edat $A_j = (a_{j-1}, a_j]$ per $j = 1, .., l$, on $a_0 = 0$.

Per a cada interval $(a_{j-1}, a_j]$, s'han obtingut el nombre de dones incidents, n_I, i el nombre de dones mortes per altres causes, n_M, de la seguent manera

$$n_{Ij} \sim Poisson(\lambda = N_{j-1} I_j)$$
$$n_{Mj} \sim Poisson(\lambda = N_{j-1} M_j)$$

S'ha emprat la recurrència $N_j = N_{j-1} - n_{Ij} - n_{Mj}$ per $j = 1, .., l$ amb $N_0 = N = 100000$ per a calcular els nombre de dones vives al final de l'interval A_j. S'han obtingt el nombre de dones incidents al llarg del seguiment de la cohort, n, a partir de la suma dels n_I,

$$n = \sum_{\forall j} n_{Ij}.$$

Per al càlcul de les edats d'incidència, X_C, s'ha considerat que les dones es posen incidents de manera uniforme durant l'interval. Per a cada interval $(a_{j-1}, a_j]$

$$X_{C_k} \sim U(a_{j-1}, a_j), \quad k = 1, .., n_{Ij}$$

2.4. Generació dels temps de supervivència.

2.4.1. *Temps de supervivència en situació de riscos competitius.* S'han simulat els temps de supervivència de les n dones incidents a partir del mètode descrit per Beyersmann per a generar dades de supervivènica en situació de riscos competitius [**25**]. La supervivència en situació de riscos competitius, en aquest cas amb dues causes, està definida per les següents expressions

- $X_t \in \{0, 1, 2\}$ és un procés estocàstic amb diferents causes de finalització. Amb $P(X_0 = 0) = 1$.
- El temps de supervivència $T = inf\{t | X_t \neq 0\}$, per a la causa de fallida $X_T \in \{1, 2\}$.
- Les funcions de risc específiques per causa, λ_{0i}, venen determinades completament pel comportament dels riscos competitius,
$$\lambda_{0i}(t)dt = P(T \in dt, X_T = i | T \geq t), \ i = 1, 2.$$
- Funció de supervivència
$$P(T > t) = \exp\left(- \int_0^t \lambda_{01}(t) + \lambda_{02}(t) du \right).$$

- Funció d'incidència acumulada

$$P(T \leq t, X_T = i) = \int_0^t P(T > u-)\lambda_{0i}(t)du, \ i = 1, 2.$$

Els autors recomanen la següent estratègia de simulació.

(1) Especificar les funcions de risc específiques per a cada causa $\lambda_{01}(t)$ i $\lambda_{02}(t)$

(2) Simular temps de supervivència T a partir de la funció de risc per a totes les causes $\lambda_{0.} = \lambda_{01}(t) + \lambda_{02}(t)$.

(3) Decidir per la causa $X_T = 1$ amb una probabilitat binomial $\lambda_{01}(T)/\lambda_{0.}(T)$.

(4) Generar temps de censura C.

En el nostre cas estan definides les variable aleatòries $X(> 0)$ com l'edat de la dona i $T(> 0)$ com el temps de supervivència des del diagnòstic fins a la mort per càncer de mama o altres causes. S'ha obtingut la funció de risc específica de morir per altres causes, $\lambda_{AC}(t)$, a partir de l'equació 39 i s'ha calculat la funció de risc específica de morir de càncer de mama, $\lambda_{CM}(t)$, a partir de la relació

$$\lambda_{CM}(t) = \frac{f_{CM}(t)}{S_{CM}(t)} = \frac{f_{CM}(t)}{\int_t^\infty f_{CM}(u)\, du}$$

S'ha generat temps de supervivència a partir de la funció de supervivència global $S(t)$ (eq. 42) mitjançant el mètode de la transformada inversa, $T = S^{-1}(u)$ on $u \sim U(0, 1)$.

S'ha assignat la causa de mort de càncer de mama per a una dona incident $C = 1$ mitjançant una variable aleatòria Bernoulli amb probabilitat d'èxit

$$p = \frac{\lambda_{CM}(T)}{\lambda(T)} = \frac{\lambda_{CM}(T)}{\lambda_{CM}(T) + \lambda_{AC}(X + T)}$$

Per aquesta simulació no cal generar cap temps de censura, ja que no hi ha una data de final de seguiment de l'estudi.

Atès que cada dona té associada una funció de risc de mort per altres causes, $\lambda_{AC}(t)$, que depèn de l'edat d'incidència del càncer de mama de la dona, s'estan generant temps de supervivència que provenen de funcions de supervivència diferents. Per tal de tenir una funció de supervivència que englobi totes aquestes funcions, s'ha estimat una funció de supervivència a partir de l'estimador de Nelson-Aalen per a la funció de risc [**22**].

S'ha simulat una mostra de temps $T_1, ..., T_n$ i un indicador de censura $\delta_i = 1$ per a tots els casos, ja que s'han generat per a totes les dones temps fins a la mort i no hi ha censura. S'han considerat els estadístics d'ordre $T_{(1)} < T_{(2)} < ... < T_{(n)}$ dels

temps observats i sigui $\delta_{(i)}$ el valor de δ associat a $T_{(i)}$. La funció de risc (instantani) al moment $T_{(i)}$ s'ha estimat a partir del nombre d'esdeveniments que s'han produït a $T_{(i)}$ dividit pel nombre d'individus a risc a $T_{(i)}$, és a dir $\hat{\lambda}(t) = d_i/n_i$

En general, l'estimador de Nelson-Aalen per a la funció de risc és:

$$\hat{\lambda}_{NA}(t) = \begin{cases} 0 & t \neq T_{(i)} \\ \frac{d_i}{n_i} & t = T_{(i)} \end{cases} \tag{49}$$

Com la funció de risc acumulada està definida $\Lambda(t) = \int_0^t \lambda(u)du$, l'estimador de Nelson-Aalen per a la mateixa és

$$\hat{\Lambda}_{NA}(t) = \begin{cases} 0 & t \leq T_{(1)} \\ \sum_{i:T_{(i)} \leq t} \frac{d_i}{n_i} & t \geq T_{(1)} \end{cases} \tag{50}$$

A partir d'aquest estimador i pel fet que $S(t) = \exp\{-\Lambda(t)\}$, es pot definir un estimador per a la supervivència com

$$\hat{S}_{NA}(t) = \exp\{-\hat{\Lambda}_{NA}(t)\} \tag{51}$$

Com que es disposen de les funcions de supervivència de càncer de mama per diferents grups d'edat d'incidència, s'ha estimat la funció de supervivència total estratificada per a tots aquests grups. D'aquesta manera s'han obtingut les estimacions de $\hat{S}_{25-39}(t)$, $\hat{S}_{40-49}(t)$, $\hat{S}_{50-59}(t)$, $\hat{S}_{60-69}(t)$ i $\hat{S}_{70-85}(t)$. S'han utilitzat splines cúbics per tal d'allisar aquestes estimacions i obtenir funcions de supervivència contínues. S'han utilitzat aquestes funcions per a generar els temps de supervivència clínics, T_C, mitjançant la còpula de Clayton.

2.4.2. *Generació de* (T_S, T_C). S'ha generat una mostra bivariant de mida n a partir d'un model de còpula de Clayton per a tot el pla (T_S, T_C) mitjançant el mètode Inverse Probability Method (Trivedi and Zimmer, 2007). Per a cada grup d'edat, s'ha establert $v_1 = \hat{S}(t)$, ja que es pot considerar la funció de supervivència com una variable aleatòria que té una distribució uniforme a l'interval $(0, 1)$. Al mateix temps, s'ha obtingut v_2 a partir d'una variable aleatòria $u \sim U(0, 1)$. S'assigna $u_1 = v_1$ i

$$u_2 = \frac{\partial C_\alpha(u_1, u_2)}{\partial u_1} = \left(v_1^{1-\alpha}\left(v_2^{(1-\alpha)/\alpha} - 1\right) + 1\right)^{1/(1-\alpha)}.$$

Llavors, s'obté $T_C = S_C^{-1}(u_1)$ i $T_S = S_S^{-1}(u_2)$, on la supervivència del temps clínic és l'estimació no paramètrica i la supervivència del temps de sojorn és una distribució exponencial de paràmetre $m = 1/\lambda$ dependent de l'edat d'entrada a l'estat pre-clínic.

S'han considerat diferents valors del paràmetre de dependència α per tal de tractar el fet que la durada del temps de sojorn de les dones mortes per altres causes no necessàriament ha d'estar correlacionat positivament amb la durada de la supervivència. És per això que s'ha assumit una correlació més forta en els grups de dones menors de 70 anys, on el risc de morir per altres causes és molt menor. En canvi, per al grup de dones diagnosticades amb més de 70 anys, s'ha considerat una correlació més baixa. En particular, $\alpha \in \{3, 3/2\}$ respectivament. Sota el model de còpula de Clayton, aquests valors representen uns valors de la tau de Kendall de $\tau_K \in \{0.5, 0.2\}$.

Per altra banda, quan es generen els temps de sojorn, T_S, encara no es té l'edat d'entrada a l'estat pre-clínic, X_S, que és la que determina el valor del paràmetre de l'exponencial segons l'expressió 34 i $\lambda = 1/m$. És per això que s'ha considerat l'edat d'incidència de cada dona, X_C, sota la suposició que l'error comès no és gran i només en uns quants casos.

A partir dels temps (T_S, T_C) i l'edat d'incidència, X_C, s'ha obtingut l'edat d'entrada a pre-clínic $X_S = X_C - T_S$ i l'edat de mort de la persona $X_D = X_C + T_C$ que s'ha classificat en mort per càncer de mama, D_{CM}, o mort per altres causes, D_{AC}, segons la variable causa de mort C.

2.4.3. *Generació d'X_L i de T_L.* Per tal de generar l'edat en el moment del diagnòstic precoç, X_L, i la durada del *lead-time*, T_L, s'ha definit la variable edat de la dona a la r-èssima mamografia, X_{Er}.

En un principi, s'ha considerat que una dona pot ser detectada per cribratge si $X_S \leq X_{Er} \leq X_C$ per a qualsevol mamografia r. S'han considerat els valors de la sensibilitat de la mamografia descrits a l'equació 35. Per a decidir si el resultat de la r-èssima mamografia ha estat positiu o negatiu s'ha utilitzat una variable aleatòria Bernoulli amb probabilitat d'èxit

$$p = \beta(X_{Er}).$$

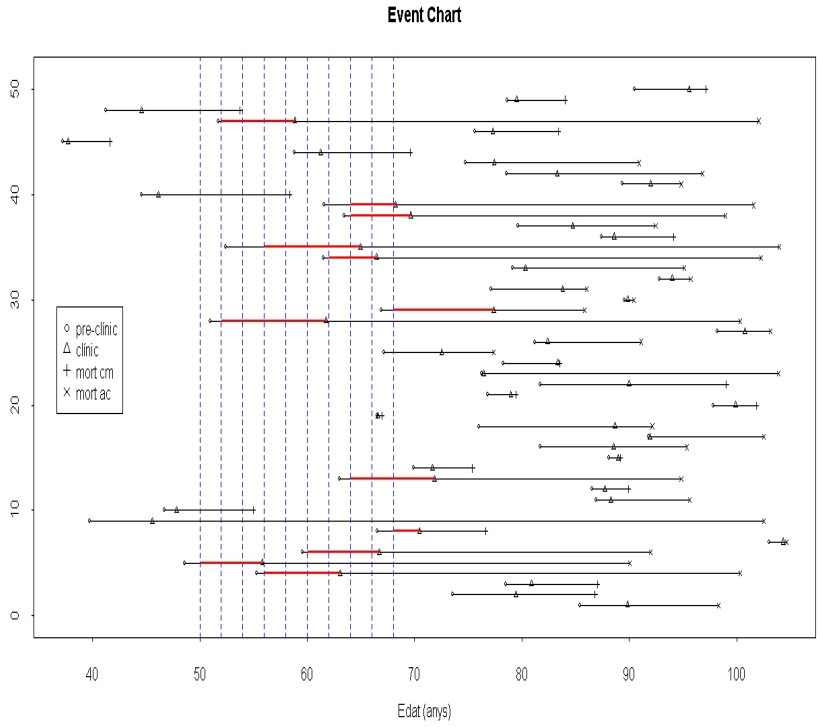
Quan la primera mamografia positiva es produeix a r, s'assigna $X_L = X_{Er}$ i es calcula $T_L = X_C - X_L$ per a les dones que han estat detectades per cribratge.

2.4.4. *Implementació de la simulació i representació de les dades.* La implementació dels mètodes de simulació ha estat duta a terme mitjançant el paquet estadístic R. El codi de la simulació està documentat en l'apèndix A.

Per tal de visualitzar el tipus de dades que s'obtenen amb aquest estudi de simulació, a la Figura 4.5 es presenta un gràfic d'una mostra aleatòria de 50 dones incidents de càncer de mama. Al llarg del temps estan situades les diferents edats generades, així com la durada de la fase pre-clínica i clínica per a cada dona.

Les línies verticals discontínues representen les edats on es duen a terme els exàmens de detecció precoç. Si una dona és detectada mitjançant la mamografia, l'edat del

FIGURA 4.5. Representació gràfica de la simulació del programa de detecció precoç



diagnòstic és avançada pel *lead-time*, en vermell. Es pot observar com, a mesura que augmenta l'edat, augmenta la incidència del càncer de mama. Les dones incidents en edats més grans moren majoritàriament d'altres causes, al contrari que les dones diagnosticades més joves, que moren principalment de càncer de mama.

2.5. Criteris d'avaluació. Per a l'avaluació dels resultats de la simulació s'ha proposat l'estimació de diferentes paràmetres d'interès com el temps mig de sojorn m, el lead-time esperat L, la incidència acumulada al llarg del seguiment de la cohort CI i la proporció de dones mortes per càncer de mama p_{CM} tant per a dones detectades precoçment com simptomàticament. S'ha calculat la sensibilitat del programa de cribratge a partir del nombre de casos detectats per examen i el nombre de casos d'interval. Es defineix la sensibilitat del programam com el quocient del nombre de casos detectats per examen entre el nombre de casos totals diagnosticats dins la finestra d'edats de cribratge.

Per als citats paràmetres, s'ha calculat el vector

$$\hat{\theta}^b = \left(\hat{\theta}_1^b, \hat{\theta}_2^b, \hat{\theta}_3^b, \hat{\theta}_4^b\right)^t$$

on $b = 1, ..., B$. Per a cada escenari, s'ha calculat la mitjana

$$\bar{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^b.$$

Els resultats obtinguts es compararan amb resultats publicats a la literatura.

Capítol 5
Resultats

1. Anàlisi del temps de supervivència de les dones diagnosticades de càncer de mama

En aquesta secció es mostren els resultats obtinguts a partir de les dades dels Registres de Càncer de Girona i Tarragona, i les dades del programa de detecció precoç de l'Hospital del Mar. S'ha fet una anàlisi descriptiva i una anàlisi de supervivència específica i relativa del càncer de mama.

1.1. Anàlisi descriptiva amb dades dels Registres de Càncer. La Taula 5.1 conté els principals resultats. L'edat mitjana dels casos en el moment del diagnòstic va ser de 61.37 anys (desviació estàndard de 14.99). Per grups d'edat, el que té més casos és el de més de 70 anys. Del total de dones diagnosticades de càncer de mama entre 2000 i 2006 un 16.8% van ser detectades mitjançant exàmens de cribratge. La mortalitat de totes les dones diagnosticades va ser del 23.8% a la fi del seguiment. El 14.7% de les dones diagnosticades van morir de càncer de mama.

Es van trobar diferències estadísticament significatives segons el tipus de detecció (cribratge, no cribratge). A la Taula 5.2 es comparen les característiques de les dones en els dos grups.

Es pot observar com les dones detectades precoçment són majoritàriament diagnosticades entre els 50 i els 70 anys, edats que corresponen a la població diana dels programes de detecció precoç. En canvi, les dones diagnosticades quan apareixen els símptomes de la malaltia tenen una distribució d'edat prou homogènia entre els grups d'edat, encara que hi ha més casos diagnosticats a partir dels 70 anys i menys casos en el grup de menys de 40 anys. S'aprecia força diferència entre els dos grups en termes de mortalitat global durant el temps de seguiment, havent-hi un 5.5% de dones mortes en el grup de dones detectades precoçment i un 27.4% en el grup simptomàtic. També hi ha diferències estadísticament significatives en la causa de mort. Van morir per càncer de mama el 2.8% de les dones detectades per cribratge, i el 17% de les dones no detectades per cribratge.

TAULA 5.1. Característiques de les dones diagnosticades de càncer de mama

Variable		**n**	%
Edat (anys)	≤ 40	316	8.0
	41-50	765	19.3
	51-60	883	22.2
	61-70	781	19.7
	> 70	1224	30.8
	Total	3969	100.0
Tipus de	No cribratge	3303	83.2
detecció	Cribratge	668	16.8
	Total	3971	100.0
Estat final	Viva	3028	76.2
del seguiment	Morta	944	23.8
	Total	3972	100.0
Mort CM	No	3389	85.3
	Sí	583	14.7
	Total	3972	100.0

TAULA 5.2. Característiques de les dones diagnosticades de càncer de mama segons tipus de detecció

Variable		$\mathbf{n}_{\text{nocribratge}}$	$\%_{\text{nocribratge}}$	$\mathbf{n}_{\text{cribratge}}$	$\%_{\text{cribratge}}$
Edat (anys)	≤ 40	316	9.6	0	0.0
	41-50	731	22.1	34	5.1
	51-60	517	15.7	366	54.8
	61-70	514	15.6	267	40.0
	> 70	1222	37.0	1	0.1
$p = 0.0005$	Total	3300	100.0	668	100.0
Estat final	Viva	2397	72.6	631	94.5
del seguiment	Morta	906	27.4	37	5.5
$p < 0.0001$	Total	3303	100.0	668	100.0
Mort CM	No	2739	82.9	649	97.2
	Sí	564	17.1	19	2.8
$p < 0.0001$	Total	3303	100.0	668	100.0

1.2. Anàlisi descriptiva amb dades del programa de detecció precoç de càncer de mama de l'Hospital del Mar. La Taula 5.3 conté els principals resultats. L'edat mitjana dels casos en el moment del diagnòstic va ser de 62.85 anys (desviació estàndard de 13.3). La majoria de les dones diagnosticades tenen més de 50 anys. Un 26.9% de les dones van ser diagnosticades en els exàmens de cribratge, un 10% més que als registres. El 80.3% dels tumors han estat diagnosticats en els estadis I i II, on el pronòstic és més favorable. La mortalitat de les dones diagnosticades en exàmens de cribratge va ser del 25.5% a la fi del seguiment. La mortalitat específica per càncer de mama va ser del 6.6%.

La Taula 5.4 presenta les característiques de les dones estudiades segons el tipus de detecció. Les dones diagnosticades en exàmens de cribratge són més joves, tenen una distribució d'estadis de la malaltia més favorable i una mortalitat global i

TAULA 5.3. Característiques de les dones diagnosticades de càncer de mama

Variable		**n**	%
Edat	≤ 40	75	4.4
	41-50	248	14.6
	51-60	416	24.4
	61-70	464	27.2
	> 70	501	29.4
	Total	1704	100.0
Tipus de detecció	No cribratge	1246	73.1
	Cribratge	458	26.9
	Total	1704	100.0
Estadi	Estadi 1	568	38.6
	Estadi 2	614	41.7
	Estadi 3	216	14.7
	Estadi 4	74	5.0
	Total	1472	100.0
Estat final del seguiment	Viva	1269	74.5
	Morta	435	25.5
	Total	1704	100.0
Mort CM	No	1591	93.4
	Sí	113	6.6
	Total	1704	100.0

per càncer de mama més baixes que les dones diagnosticades quan apareixen els símptomes.

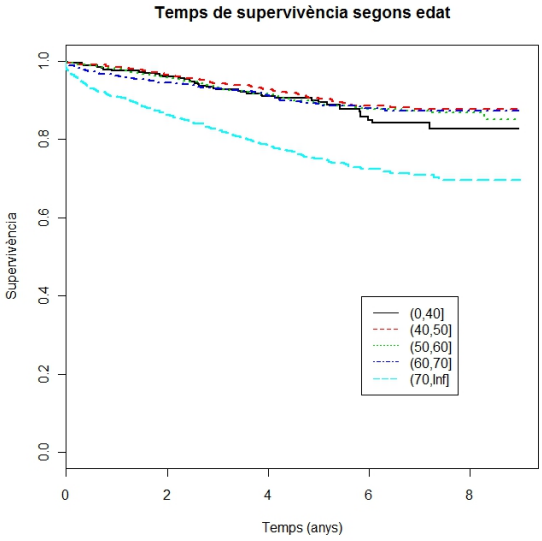
TAULA 5.4. Característiques de les dones diagnosticades de càncer de mama segons cribratge

Variable		$\mathbf{n}_{\text{nocribratge}}$	$\%_{\text{nocribratge}}$	$\mathbf{n}_{\text{cribratge}}$	$\%_{\text{cribratge}}$
Edat	≤ 40	74	5.9	1	0.2
	41-50	217	17.4	31	6.8
	51-60	208	16.7	208	45.4
	61-70	247	19.8	217	47.4
	> 70	500	40.1	1	0.2
$p = 0.0005$	Total	1246	100.0	458	100.0
Estadi	Estadi 1	316	30.1	252	59.4
	Estadi 2	478	45.6	136	32.1
	Estadi 3	184	17.6	32	7.5
	Estadi 4	70	6.7	4	0.9
$p = 0.0005$	Total	1048	100.0	424	100.0
Estat UC	Viva	853	68.5	416	90.8
	Morta	393	31.5	42	9.2
$p < 0.0001$	Total	1246	100.0	458	100.0
Mort CM	No	1139	91.4	452	98.7
	Sí	107	8.6	6	1.3
$p < 0.0001$	Total	1246	100.0	458	100.0

1.3. Anàlisi de la supervivència amb dades dels Registres de Càncer. S'han estimat les funcions de supervivència específiques de mort per càncer de

mama estratificades per edat i tipus de detecció mitjançant l'estimador de Kaplan-Meier [**22**].

FIGURA 5.1. Funció de supervivència del càncer de mama estratificada per edat al diagnòstic



A la Figura 5.1 es pot observar com el temps de supervivència de càncer de mama és semblant en tots els grups. El grup de més de 70 anys té una supervivència inferior comparada amb els altres grups, però cal destacar que hi ha un percentatge gran de censurats i que el risc de morir per altres causes és elevat.

A la Figura 5.2 es pot apreciar que existeixen diferències significatives en la funció de supervivència acumulada segons el tipus de detecció. El grup de dones detectades precoçment té una corba de supervivència pràcticament plana al llarg del temps.

A la Figura 5.3, on es presenta la supervivència relativa, s'observen resultats similars. Hi ha un increment de la supervivència al cap de 8 anys, ja que el nombre de morts en aquest temps és menor que l'esperat. Aquest fet s'atribueixt als controls mèdics de les dones diagnosticades de càncer de mama durant el seguiment de la malaltia [**26**].

1.4. Anàlisi de la supervivència amb dades del programa de detecció precoç de càncer de mama de l'Hospital del Mar. S'han estimat les funcions de supervivència específiques de mort per càncer de mama estratificades per edat, estadi de la malaltia i tipus de detecció mitjançant l'estimador de Kaplan-Meier [**22**].

A la Figura 5.4 es pot observar que el grup d'edat de més de 70 anys té una supervivència més desfavorable respecte als altres grups.

FIGURA 5.2. Funció de supervivència del càncer de mama estratificada per tipus de detecció

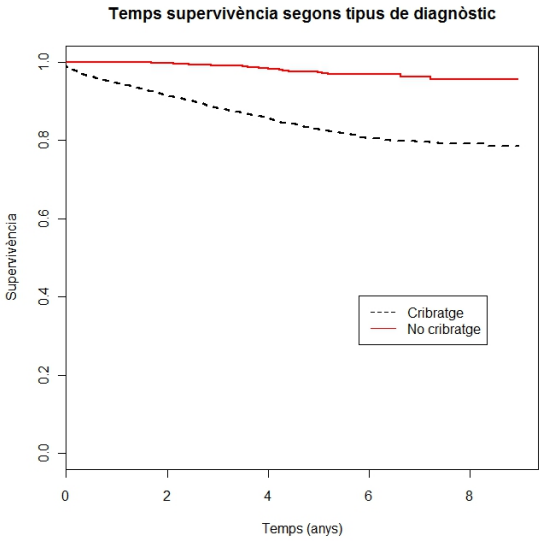
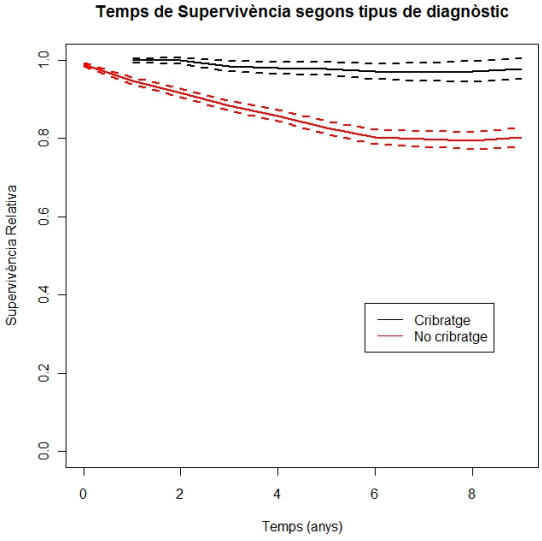


FIGURA 5.3. Funció de supervivència relativa del càncer de mama estratificada per tipus de detecció



A la Figura 5.5 es pot observar que la supervivència empitjora a mesura que avança l'estadi de la malaltia. Destaca el grup de dones diagnosticades amb Estadi IV, equivalent a malaltia metastàsica.

FIGURA 5.4. Funció de supervivència del càncer de mama estratificada per edat al diagnòstic

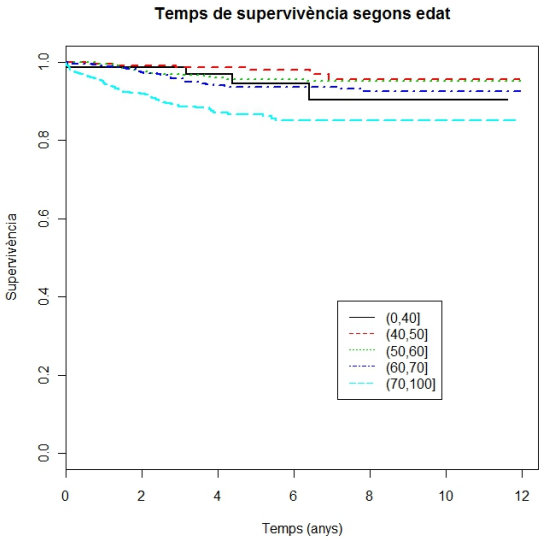
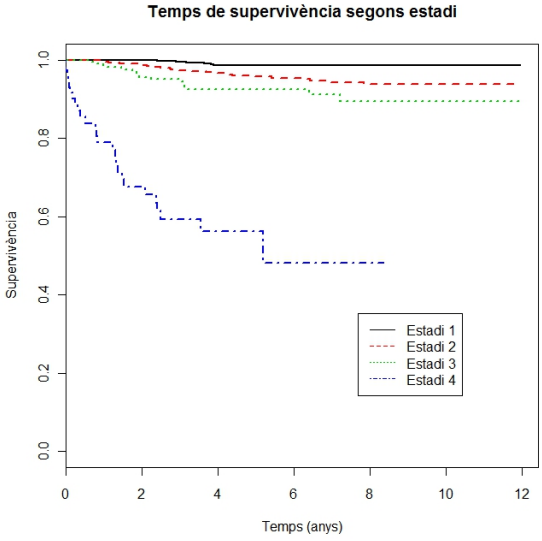


FIGURA 5.5. Funció de supervivència del càncer de mama estratificada per estadi



La Figura 5.6 presenta les corbes de supervivència segons tipus de detecció. La corba de supervivència de les dones detectades precoçment és pràcticament plana al llarg del temps de seguiment.

FIGURA 5.6. Funció de supervivència del càncer de mama estratificada per tipus de detecció

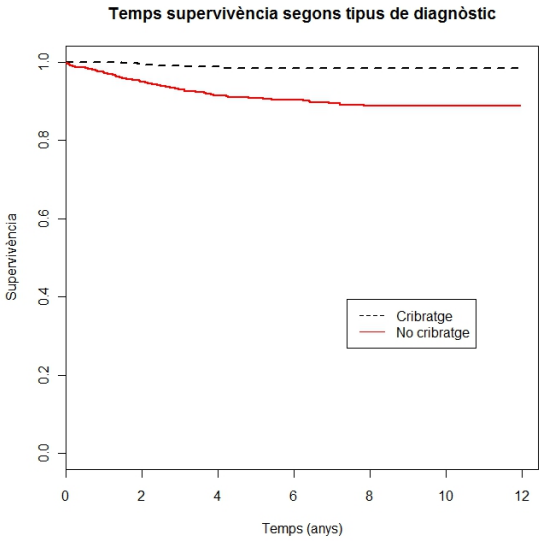
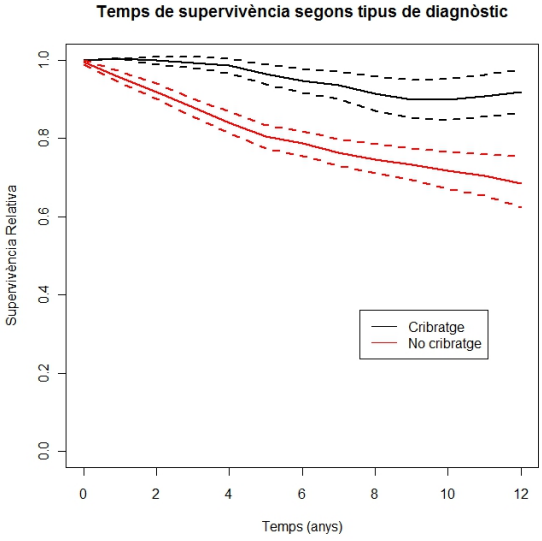


FIGURA 5.7. Funció de supervivència relativa del càncer de mama estratificada per tipus de detecció



A la Figura 5.7 presenta la supervivència relativa segons tipus de detecció. De nou, s'observa un increment de la supervivència al final del seguiment, que com ja s'ha dit a la secció anterior podria estar causat pel control mèdic de les pacients.

2. Resultats de la simulació

En aquesta secció es mostren els resultats obtinguts amb la simulació dels quatre escenaris de cribratge que s'han considerat. S'han generat 100 conjunts de 100.000 dones que reprodueixen la història natural de la malaltia segons els patrons d'incidència del càncer de mama i mortalitat de Catalunya. Els quatre escenaris de cribratge, que varien segons l'edat d'inici i la periodicitat dels exàmens, s'han aplicat a les dones generades per a obtenir estimacions dels paràmetres d'interès.

A la secció 2.1 es presenta gràficament un exemple de cada escenari analitzat. La secció 2.2 presenta l'estimació dels paràmetres d'interès a partir dels conjunts de dades generats i mostra els resultats de supervivència.

2.1. Exemples dels escenaris de la simulació. A les Figures 5.8-5.11 es pot observar la història natural de la malaltia d'una mostra de 50 dones incidents. Les línies verticals corresponen als exàmens de cribratge. En funció de la sensibilitat de la prova diagnòstica es produeïx la detecció precoç que determina el temps d'avenç del diagnòstic o *lead-time* representat en color vermell. Podem observar com, a mesura que augmenta l'edat augmenta la incidència del càncer de mama. Les dones incidents més grans moren majoritàriament d'altres causes, al contrari que les dones diagnosticades en edats més joves, que moren principalment de càncer de mama.

FIGURA 5.8. Història natural de la malaltia i lead-time d'una mostra de dones simulades. Exàmens biennals i edat d'inici 40 anys

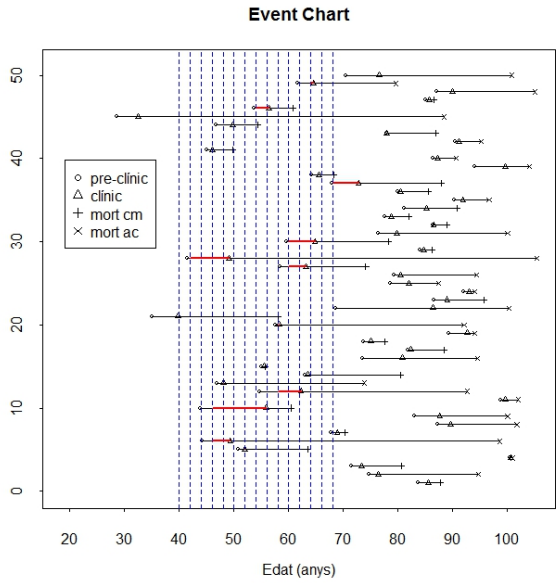


FIGURA 5.9. Història natural de la malaltia i lead-time d'una mostra de dones simulades. Exàmens anuals i edat d'inici 40 anys

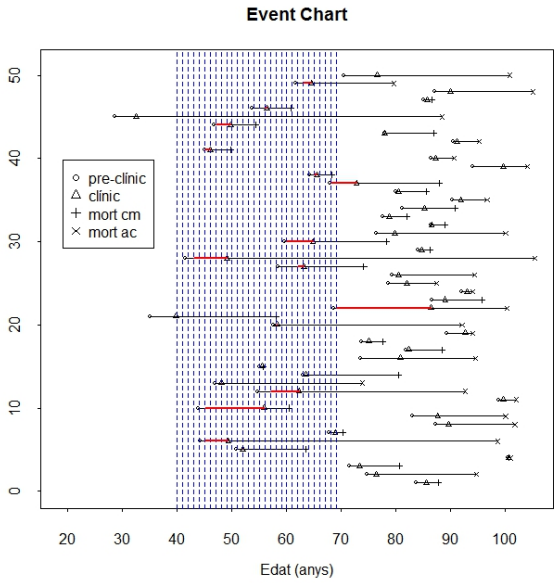


FIGURA 5.10. Història natural de la malaltia i lead-time d'una mostra de dones simulades. Exàmens biennals i edat d'inici 50 anys

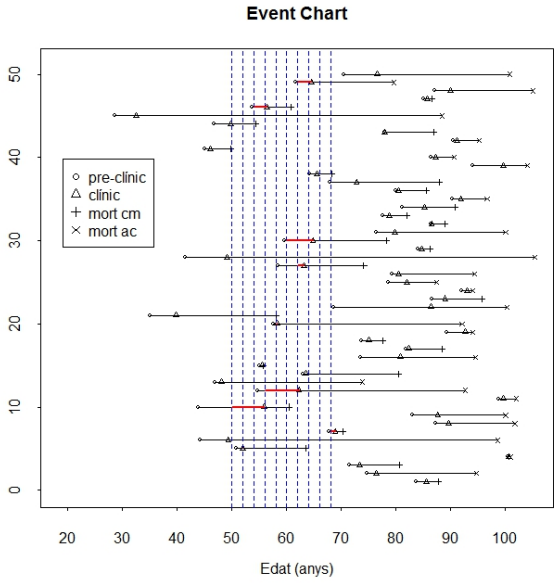
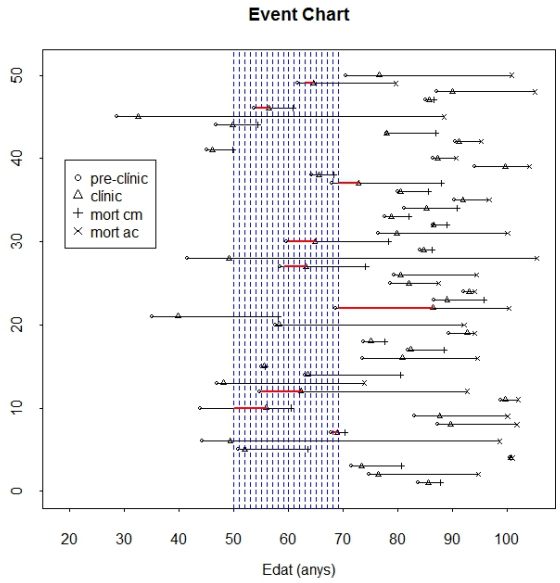


FIGURA 5.11. Història natural de la malaltia i lead-time d'una mostra de dones simulades. Exàmens anuals i edat d'inici 50 anys

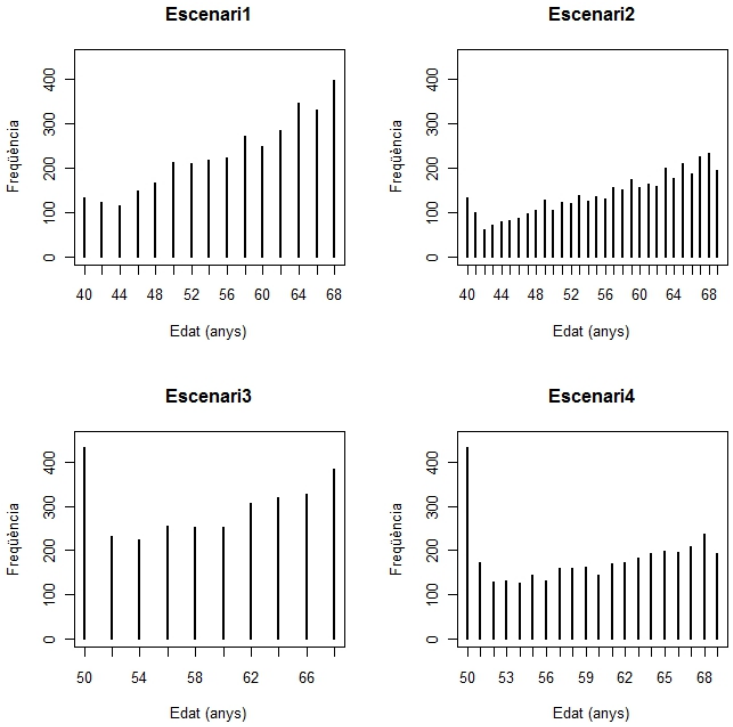


2.2. Estimació dels paràmetres de la simulació. A la Figura 5.12 es presenta el nombre de casos detectats a cada examen segons els diferents escenaris de cribratge. S'observa un nombre més elevat de casos detectats en el primer examen de cribratge. Corresponen a l'anomenat cribratge de prevalença.

A la taula 5.5 es mostren les estimacions dels paràmetres obtingudes fent les mitjanes dels resultats de la simulació, per a cada escenari. Per als casos incidents s'ha calculat la mitjana del temps de sojorn i el *lead-time* esperat. S'ha obtingut la incidència acumulada al llarg del seguiment de la cohort i el percentatge de morts de càncer de mama. S'ha calculat la sensibilitat de cada programa de cribratge a partir dels càncers detectats precoçment i els càncers d'interval, durant la durada del programa. Per aquests dos tipus de càncers s'ha calculat el temps de sojorn corresponent.

D'acord amb el disseny de la simulació, el nombre de casos incidents de càncer de mama, la mitjana del temps de sojorn en estat pre-clínic i la mortalitat per càncer de mama no varien pels quatre escenaris. Cal esmentar que en aquest estudi no s'ha tingut en compte el benefici del cribratge en termes de reducció de la mortalitat per càncer de mama. La mitjana del temps de sojorn (3.91 anys) es troba entre els límits esperats segons les assumpcions de l'estudi de simulació. La majoria de dones incidents tenen més de 50 anys en el moment de la detecció, edat a partir de la qual la mitjana del temps de sojorn introduït en la simulació és 4 anys. En els casos detectats per exàmens, el *lead-time* esperat varia entre 4.69 i 4.83 anys segons l'escenari. Augmenta lleugerament en augmentar la freqüència i disminuïr l'edat d'inici dels exàmens. Aquest temps respresenta el biaix del temps de supervivència

FIGURA 5.12. Nombre de casos detectats per examen segons programa de cribratge



causat per la detecció precoç. El nombre de casos detectats per examen i el nombre de càncers d'interval varia també segons l'escenari. A més exàmens i menor edat d'inici del cribratge es detecten més casos precoçment i sorgeixen menys càncers d'interval. Per tant la sensibilitat del programa es veu afectada per aquest fenomen. Les tres darreres línies de la taula mostren la mitjana del temps de sojorn en els casos detectats per examen, els càncers d'interval i els casos no detectats per examen, els quals incloen els càncer d'interval i els casos diagnosticats fora de la finestra d'edats de cribratge. Es pot observar l'efecte del biaix de durada (*length bias*) amb temps de sojorn considerablement superiors en els casos detectats per examen en relació a la resta. Destaca especialment el valor petit de la mitjana del temps de sojorn dels càncers d'interval, que per naturalesa són més agressius.

A la Figura 5.13 es presenta la distribució del *lead-time* segons l'escenari. S'observa el patró exponencial, resultat consistent amb les assumpcions utilitzades en la simulació (temps de sojorn exponencial).

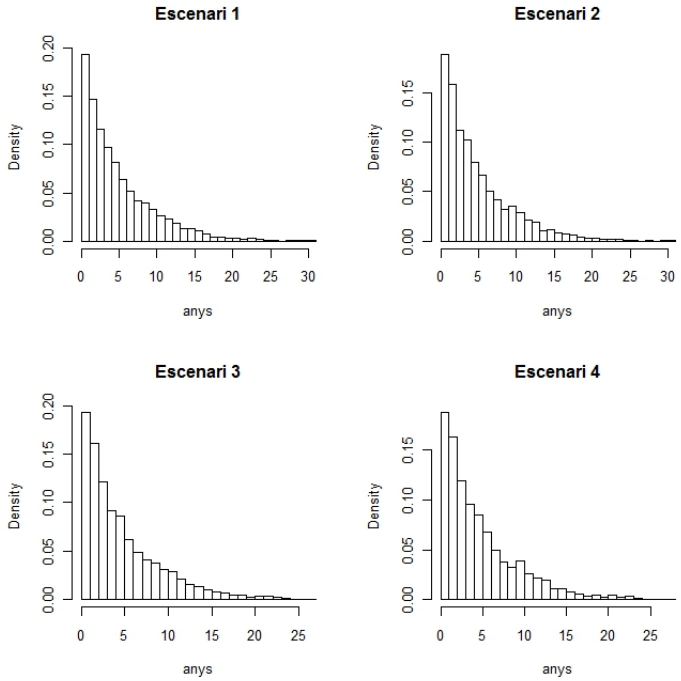
A la Figura 5.14 es mostra l'efecte de la detecció precoç en el càlcul de la funció de supervivència en els casos detectats per examen, en els quatre escenaris estudiats. La corba contínua indica la supervivència lliure del biaix de detecció precoç, calculada amb els temps de supervivència a partir de l'entrada a l'estat clínic S_c, anomenat *post-lead-time*. La corba discontinua està afectada pel biaix de detecció

TAULA 5.5. Paràmetres estimats segons els escenaris de cribratge

Mètode	**Escenari 1**[1]	**Escenari 2**[2]	**Escenari 3**[3]	**Escenari 4**[4]
Casos Incidents	11655.99	11655.99	11655.99	11655.99
Mitjana temps de sojorn	3.91	3.91	3.91	3.91
Lead-time esperat	4.83	4.79	4.71	4.69
Incidència acumulada (%)	11.66	11.66	11.66	11.66
Morts per CM (%)	50.45	50.45	50.45	50.45
Detectats per mamografia	3378.22	4112.85	2933.27	3558.35
D'interval	1180.63	728.57	912.05	569.45
Sensibilitat programa (%)	74.10	84.95	76.28	86.2
Mitjana temps sojorn D.[5]	6.28	5.68	6.54	5.95
Mitjana temps sojorn I.[5]	1.17	0.71	1.36	0.89
Mitjana temps sojorn ND.[5]	2.94	2.94	3.03	3.01

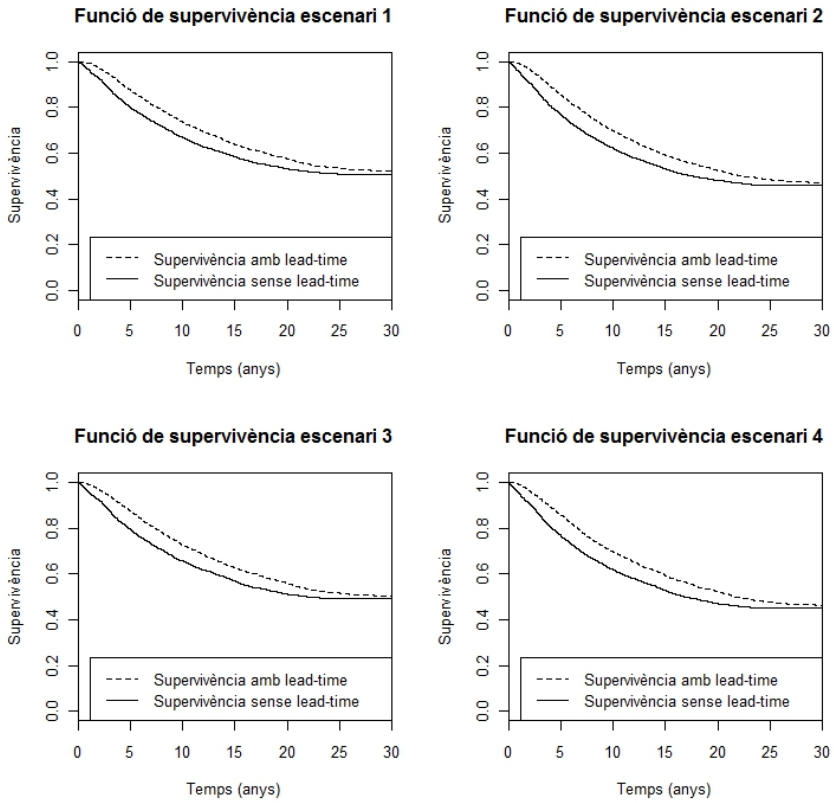
[1] Escenari 1: Exàmens biennals i edat d'inici 40 anys.
[2] Escenari 2: Exàmens anuals i edat d'inici 40 anys.
[3] Escenari 3: Exàmens biennals i edat d'inici 50 anys.
[4] Escenari 4: Exàmens anuals i edat d'inici 50 anys.
[5] D: càncers detectats precoçment, I:càncer d'interval, ND: càncers no detectats

FIGURA 5.13. Distribució lead-time per escenari



precoç. S'ha obtingut sumant el *lead-time* i el temps de supervivència *post-lead-time* per a cada individu.

FIGURA 5.14. Supervivència amb l'efecte del lead-time per escenari



3. Comparació dels mètodes estudiats

3.1. Comparació dels mètodes amb dades dels registres de Catalunya. A la taula 5.6 es presenten les dades de supervivència de les dones diagnosticade per cribratge, en format de taula de vida, corresponents a les dones diagnosticades en els registres de Girona i Tarragona.

Les dades observades han permès utilitzar un temps de seguiment de 9 anys per estimar la funció de supervivència des del diagnòstic. Les estimacions de supervivència són elevades a causa del nombre reduït de defuncions per càncer de mama (Figura 5.15).

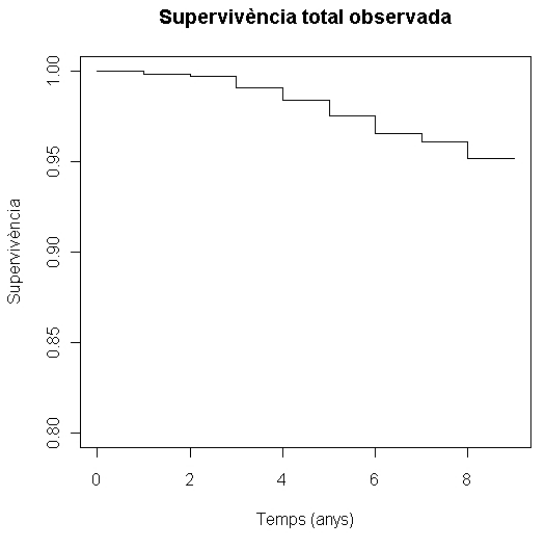
A partir de la supervivència observada és possible estimar el *post-lead-time* X, mitjançant els mètodes de Walter & Stitt, Xu i Prorok i Xu et al. Una assumpció necessària en tots aquests models és que el temps de sojorn T és exponencial amb mitjana $m = 1/\lambda$. S'han assumit $\lambda = 0.25$.

A la Taula 5.7 es presenten els resultats. Es pot comprovar que tots els mètodes redueixen les estimacions de supervivència en relació a la supervivència observada, afectada pel biaix *lead-time*. El mètode de Walter & Stitt obté estimacions més

TAULA 5.6. Supervivència observada per a les dones diagnosticades per cribratge

I_i	n_i	l_i	d_i	$\hat{S}_Z(z_i)$
0-1	668	3	1	0.9984996
1-2	664	2	1	0.9969936
2-3	661	8	4	0.9909236
3-4	649	140	4	0.9840779
4-5	505	134	4	0.9750908
5-6	367	114	3	0.9656545
6-7	250	95	1	0.9608858
7-8	154	104	1	0.9514654
8-9	49	49	0	0.9514654

FIGURA 5.15. Funció de supervivència observada per les dones detectades per cribratge



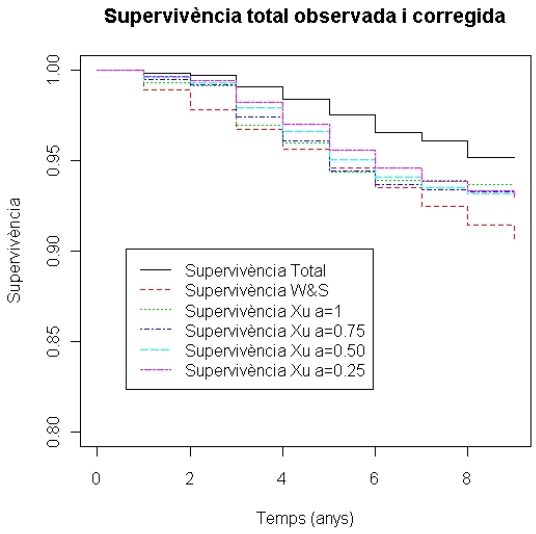
baixes de la supervivència, probablement per l'assumpció de funció de risc constant al llarg del temps. El valor de l'estimador màxim versemblant pel paràmetre de la funció de risc, θ, és 0.01115. Les estimacions obtingude aplicant els mètodes de Xu & Prorok i de Xu et al. arriben a resultats similars. No obstant, quan s'ha considerat una correlació més elevada entre el *lead-time* i el *post-lead-time*, cas amb $c = 0.25$, s'han obtingut estimacions més baixes de la funció de supervivència.

La Figura 5.16 mostra les funcions de supervivència observada i corregides pel *lead-time*. Sembla que les estimacions obtingudes pel mètode de Xu són pràcticament equivalents, encara que com cita l'autor en el seu treball, l'assumpció d'indepèndencia podria comportar una sobrestimació del *post-lead-time*.

TAULA 5.7. Estimacions del post-lead-time per mètode

I_i	$\hat{S}_X(x_i)_{WS}$	$\hat{S}_X(x_i)_{XU}$	$\hat{S}_X(x_i)_{XU_{0.25}}$	$\hat{S}_X(x_i)_{XU_{0.50}}$	$\hat{S}_X(x_i)_{XU_{0.75}}$
0-1	0.989	0.993	0.997	0.996	0.995
1-2	0.978	0.992	0.994	0.993	0.992
2-3	0.967	0.970	0.982	0.979	0.974
3-4	0.956	0.960	0.970	0.966	0.961
4-5	0.946	0.943	0.956	0.951	0.944
5-6	0.935	0.939	0.946	0.941	0.937
6-7	0.925	0.939	0.939	0.935	0.934
7-8	0.915	0.937	0.933	0.932	0.933
8-9	0.904	0.937	0.930	0.930	0.932

FIGURA 5.16. Funció de supervivència observada i del *post-lead-time* estimada per cada mètode



3.2. Comparació dels mètodes amb dades del programa de detecció precoç de càncer de mama de l'Hospital del Mar. La taula 5.8 mostra les estimacions de supervivència a partir de les dades observades.

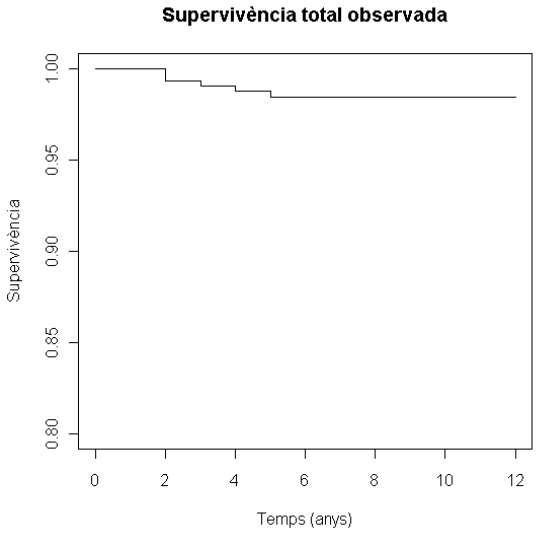
S'han utilitzat 12 anys de seguiment per estimar la funció de supervivència des del diagnòstic. Per a les dones detectades per cribratge s'han obtingut unes estimacions de la supervivència força altes a causa del baix nombre de casos de mort per càncer de mama. No hi ha cap defunció a partir dels 5 anys de seguiment (Figura 5.17).

A partir de la supervivència observada s'ha estimat el *post-lead-time* X, mitjançant els mètodes de Walter& Stitt, Xu & Prorok i Xu et al. S'ha assumit que el temps de sojorn T és exponencial de mitjana $m = 1/\lambda = 1/0.25 = 4$. A la Taula 5.9 es mostren els resultats obtinguts.

TAULA 5.8. Supervivència observada per a les dones diagnosticades per cribratge

I_i	n_i	l_i	d_i	$\hat{S}_Z(z_i)$
0-1	458	32	0	1.0000000
1-2	426	6	3	0.9929078
2-3	417	46	1	0.9903877
3-4	370	51	1	0.9875129
4-5	318	55	1	0.9841135
5-6	262	59	0	0.9841135
6-7	203	40	0	0.9841135
7-8	163	35	0	0.9841135
8-9	128	39	0	0.9841135
9-10	89	41	0	0.9841135
10-11	48	25	0	0.9841135
11-12	23	21	0	0.9841135

FIGURA 5.17. Funció de supervivència observada per les dones detectades per cribratge



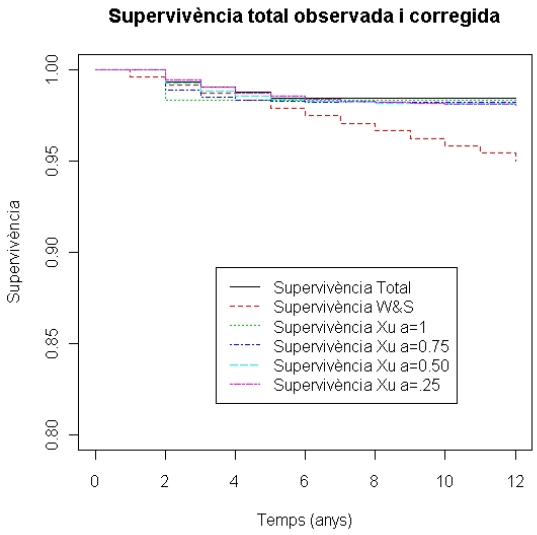
Novament, el mètode de Walter & Stitt estima una supervivència més baixa que la resta, obtinguda partir d'una distribució del *post-lead-time* exponencial de paràmetre $\hat{\theta} = 0.0043$, estimat per màxima versemblança. Les estimacions dels mètodes de Xu & Prorok i Xu et al. arriben a resultats similars entre ells, encara que per valors baixos del paràmetre c, aconsegueixen estimacions més baixes.

La Figura 5.18 mostra les funcions de supervivència observada i corregides pel *lead-time*. Es podria considerar que les estimacions del mètode de Xu & Prorok i Xu et al. són pràcticament equivalents i molt pròximes a la supervivència observada.

TAULA 5.9. Estimacions del post-lead-time per mètode

I_i	$\hat{S}_X(x_i)_{WS}$	$\hat{S}_X(x_i)_{XU}$	$\hat{S}_X(x_i)_{XU_{0.25}}$	$\hat{S}_X(x_i)_{XU_{0.50}}$	$\hat{S}_X(x_i)_{XU_{0.75}}$
0-1	0.996	1.000	1.000	1.000	1.000
1-2	0.991	0.983	0.994	0.993	0.989
2-3	0.987	0.983	0.990	0.988	0.985
3-4	0.983	0.983	0.987	0.985	0.983
4-5	0.979	0.983	0.985	0.984	0.982
5-6	0.975	0.983	0.984	0.983	0.982
6-7	0.971	0.983	0.983	0.982	0.982
7-8	0.966	0.983	0.982	0.982	0.982
8-9	0.962	0.983	0.981	0.981	0.982
9-10	0.958	0.983	0.981	0.981	0.982
10-11	0.954	0.983	0.981	0.981	0.982
11-12	0.950	0.983	0.980	0.981	0.982

FIGURA 5.18. Funció de supervivència observada i del *post-lead-time* estimada per cada mètode



En aquest cas, es tracta d'una supervivència pràcticament plana i ha estat dificil estimar l'efecte del *lead-time bias*.

3.3. Comparació dels mètodes amb dades de l'estudi de simulació. A les Figures 5.19-5.22 es mostren les diferents corbes de supervivència del post-lead-time. El mètode de Walter & Stitt no aconsegueix un ajust satisfactori del *post-lead-time*. Les estimacions dels mètodes de Xu & Prorok i Xu et al. arriben a resultats similars entre ells, encara que per valors baixos del paràmetre c, aconsegueixen estimacions més baixes.

FIGURA 5.19. Funció de supervivència clínica i estimada per cada mètode. Exàmens biennals i edat d'inici 40 anys

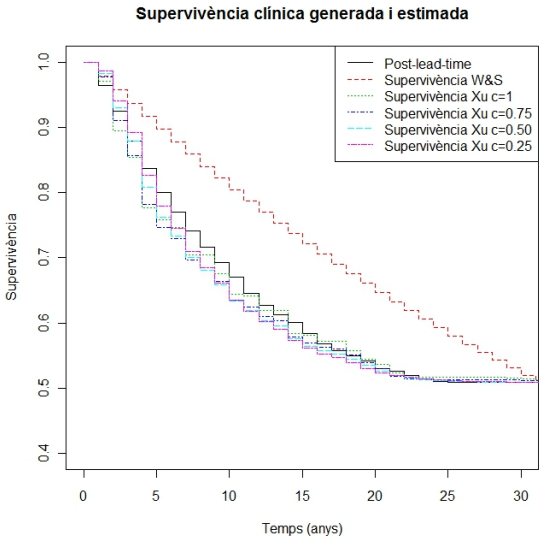
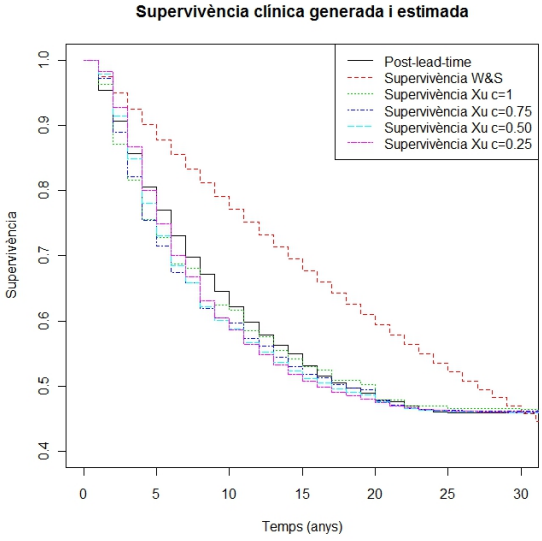


FIGURA 5.20. Funció de supervivència clínica i estimada per cada mètode. Exàmens anuals i edat d'inici 40 anys



Per tal comparar l'ajust dels mètodes a la supervivència *post-lead-time* s'ha utilitzat l'error quadràtic mig (EQM) [**27**]. La Taula 5.10 conté l'error quàdratic mig dels mètodes analitzats per als diferents escenaris. S'obtenen resultats similars en tots els mètodes, excepte en el mètode de Walter & Stitt on l'error quadràtic mig és

FIGURA 5.21. Funció de supervivència clínica i estimada per cada mètode. Exàmens biennals i edat d'inici 50 anys

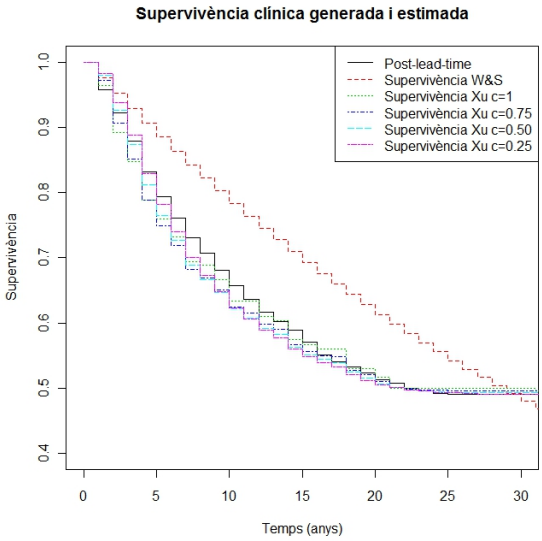
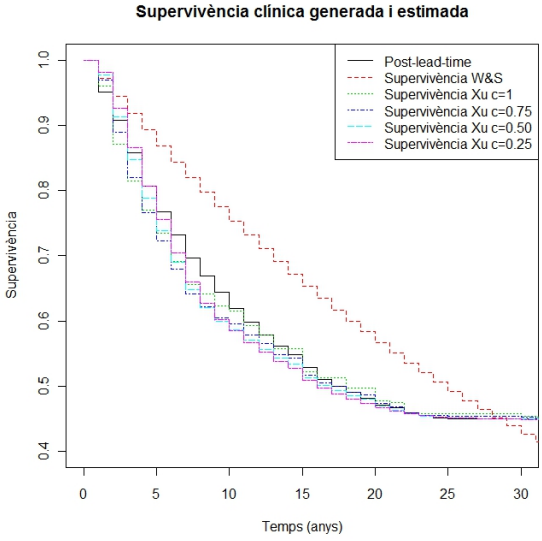


FIGURA 5.22. Funció de supervivència clínica i estimada per cada mètode. Exàmens anuals i edat d'inici 50 anys



molt elevat. En el mètode de Xu et al. s'observa un major error quadràtic mig a mesura que el paràmetre c, relacionat amb el grau de dependència entre el *lead-time* i el *post-lead-time*, augmenta. Amb l'assumpció d'independència s'obté l'error quadràtic mig més petit, excepte per a l'escenari 1.

TAULA 5.10. EQM dels mètodes per escenaris

Mètode	**Escenari 1**[1]	**Escenari 2**[2]	**Escenari 3**[3]	**Escenari 4**[4]
Walter & Stitt	5.5989	7.8635	2.9204	3.8879
Xu & Prorok	0.01529	0.0167	0.0137	0.0155
Xu et al. (c =0.25)	0.0152	0.0203	0.0150	0.0172
Xu et al. (c =0.5)	0.0181	0.0229	0.0176	0.0198
Xu et al. (c =0.75)	0.0217	0.0262	0.0204	0.0232

[1] Escenari 1: Exàmens biennals i edat d'inici 40 anys.
[2] Escenari 2: Exàmens anuals i edat d'inici 40 anys.
[3] Escenari 3: Exàmens biennals i edat d'inici 50 anys.
[4] Escenari 4: Exàmens anuals i edat d'inici 50 anys.

Capítol 6
Discussió

1. Valoració dels resultats obtinguts

En aquest treball s'ha avaluat l'impacte que els biaixos lead-time i length-time tenen sobre el càlcul de la supervivència amb diferent mètodes proposats a la literatura. S'han utilitzat dades dels registres de cancer de catalunya i d'un programa de detecció precoç. S'ha dut a terme un estudi de simulacio per superar les limitacions que tenien les dades observades. Aquest estudi de simulació ha permès avaluar amb detall i quantificar els biaixos lligats a la detecció precoç.

S'han obtingut estimacions del temps d'avenç del diagnòstic en diferents escenaris de cribratge. S'han avaluat les diferencies entre els casos detectats per examen i els casos de càncer d'interval. El temps mig d'avenç del diagnòstic en les dones diagnosticades de càncer de mama per examen de cribratge es troba al voltant dels 5 anys. Els temps de sojorn dels casos detectats per examen són superiors als de la resta de dones incidents, com era esperable a partir de les assumpcions dels models. Exceptuant el mètode de Walter & Stitt, els mètodes comparats han proporcionat resultats similars.

S'han comparat els resultats obtinguts en l'estudi de simulació amb dades de la literatura. S'ha obtingut consistència en diferents aspectes. S'ha obtingut una incidència acumulada de càncer de mama al llarg de la vida del 11.7% comparable a l'1 de cada 8 (12.5%) publicat a la literatura [**28, 29, 30**]. La mortalitat específica per càncer de mama de les dones incidents obtinguda en la simulació també és consistent amb la literatura [**31**]. Finalment la supervivència relativa de les dones amb càncer de mama mostra un comportament creixent al llarg del temps, comparable als resultats publicats [**32, 26**].

2. Limitacions

Aquest treball té les següents limitacions. 1) Les dades dels registres de càncer no van començar a recollir la variable tipus de detecció (examen de cribratge, si/no) fins l'any 2000. Aquest fet juntament amb el reduït nombre de defuncions de les dones cribrades ha limitat l'anàlisi de supervivència i ha motivat la realització d'un

estudi de simulació. 2) S'han utlitzat les taxes d'incidència de càncer de mama i de mortalitat per altres causes obtingudes en treballs previs del grup de recerca. Per a la cohort nascuda l'any 1950, les estimacions corresponents a les edats avançades podrien no ser correctes. 3) El suport de les funcions de densitat de supervivència específica de càncer de mama utilitzades en la simulació no va més enllà de 25 anys, a partir d'aquest temps s'assumeix que no hi ha risc de morir per càncer de mama. 4) S'ha assumit independència entre el risc de morir per altres causes i el càncer de mama. Caldria aprofundir en la utilització de mètodes per riscos competitius. 5) No s'ha aprofundit en la modelització de la dependència de temps de supervivència bivariants. S'han fet assumpcions sobre el paràmetre d'associació de la còpula de Clayton que caldria validar.

3. Propostes de recerca

En el futur seria interessant ampliar els següents apartats:

- Avaluar altres mètodes de correcció dels biaixos.
- Obtenir dades de supervivència de registres de càncer d'altres àrees geogràfiques i amb períodes de seguiment més llargs.
- Actualitzar les dades d'incidència de càncer de mama i mortalitat per altres causes.
- Aprofundir en l'anàlisi de riscos competitius.
- Aprofundir en els models còpula per a dades de supervivència bivariant.

4. Conclusions

S'han obtingut estimacions del temps de supervivència del càncer de mama, corregint els biaixos associats a la detecció precoç. Els resultats obtinguts mostren la importància de considerar aquest biaixos en l'avaluació correcta de la supervivència.

5. Agraïments

Voldria donar les gràcies als registres de Càncer de Girona i Tarragona i al programa de detecció precoç de càncer de mama de l'Hospital del Mar per permetrem treballar amb les seves dades. També agraïr l'ajuda mostrada pel grup GRASS en les exposicions que hem anat fent al llarg de l'any. Així també, agraïr al Dr. Xu per la seva comunicació personal que m'ha ajudat a entendre els seus mètodes. No voldria oblidar-me de l'ajuda rebuda pels membres del grup del projecte FIS (PS09/01340, IP: Montserrat Rué) i l'oportunitat que m'ha concedit per a realitzar aquest treball.

Bibliografia

[1] Ferlay J, Bray F, Pisani P, Parkin DM: **Globocan 2002: cancer, incidence, mortality and prevalence worldwide**. Tech. Rep. 5, IARC Press 2004.

[2] Cronin KA, Feuer EJ, Clarke LD, Plevritis SK: **Impact of adjuvant therapy and mammography on U.S. mortality from 1975 to 2000: comparison of mortality results from the cisnet breast cancer base case analysis**. *J Natl Cancer Inst Monogr* 2006, **(36)**:112–121.

[3] Zelen M, Feinleib M: **On the theory of screening for chronic diseases**. *Biometrika* 1969, **56**:601–614.

[4] Shen Y, Parmigiani G: **A model-based comparison of breast cancer screening strategies: mammograms and clinical breast examinations**. *Cancer Epidemiol Biomarkers Prev* 2005, **14**:529–532.

[5] Lee SJ, Zelen M: **Scheduling periodic examinations for the early detection of disease: Applications to breast cancer**. *Journal of the American Statistical Association* 1998, **93**:1271–1281.

[6] Baker SG: **Evaluating the age to begin periodic breast cancer screening using data from a few regularly scheduled screenings**. *Biometrics* 1998, **54**:1569–1578.

[7] Walter SD, Day NE: **Estimation of the duration of a pre-clinical disease state using screening data**. *Am J Epidemiol* 1983, **118**:865–886.

[8] Day NE, Walter SD: **Simplified models of screening for chronic disease: estimation procedures from mass screening programmes**. *Biometrics* 1984, **40**:1–14.

[9] Zelen M: **Forward and backward recurrence times and length biased sampling: age specific models**. *Lifetime Data Anal* 2004, **10**:325–334.

[10] Walter SD, Stitt LW: **Evaluating the survival of cancer cases detected by screening**. *Stat Med* 1987, **6**:885–900.

[11] Xu JL, Prorok PC: **Non-parametric estimation of the post-lead-time survival distribution of screen-detected cancer cases**. *Stat Med* 1995, **14**:2715–2725.

[12] Grenander U: **On the theory of mortality measurement (Part II)**. *Skandinavisk Aktuarietdskrift* 1957, **39**:125–153.

[13] Xu JL, Fagerstrom RM, Prorok PC: **Estimation of post-lead-time survival under dependence between lead-time and post-lead-time survival**. *Stat Med* 1999, **18**:155–162.

[14] Mahnken JD, Chan W, Freeman J D H, Freeman JL: **Reducing the effects of lead-time bias, length bias and over-detection in evaluating screening mammography: a censored bivariate data approach**. *Stat Methods Med Res* 2008, **17**:643–663.

[15] Duffy SW, Nagtegaal ID, Wallis M, Cafferty FH, Houssami N, Warwick J, Allgood PC, Kearins O, Tappenden N, O'Sullivan E, Lawrence G: **Correcting for lead time and length bias in estimating the effect of screen detection on cancer survival**. *Am J Epidemiol* 2008, **168**:98–104.

[16] Cleries R, Ribes J, Galvez J, Melia A, Moreno V, Bosch FX: **Automatic calculation of relative survival through the web. The WAERS project of the Catalan Institute of Oncology**]. *Gac Sanit* 2005, **19**:71–75.

[17] Martinez-Alonso M, Vilaprinyo E, Marcos-Gragera R, Rue M: **Breast cancer incidence and overdiagnosis in Catalonia (Spain)**. *Breast Cancer Res* 2010, **12**:R58.

[18] Vilaprinyo E, Gispert R, Martinez-Alonso M, Carles M, Pla R, Espinas JA, Rue M: **Competing risks to breast cancer mortality in Catalonia**. *BMC Cancer* 2008, **8**:331.

[19] Vilaprinyo E, Rue M, Marcos-Gragera R, Martinez-Alonso M: **Estimation of age- and stage-specific Catalan breast cancer survival functions using US and Catalan survival data**. *BMC Cancer* 2009, **9**:98.

[20] Lee S, Zelen M: **A stochastic model for predicting the mortality of breast cancer**. *J Natl Cancer Inst Monogr* 2006, :79–86.

[21] Elandt-Johnson R, Johnson N: *Survival models and data analysis*. New York: Wiley and sons 1980.

[22] Gomez G: **Análisis de supervivencia**. Tech. rep., Universitat Politècnica de Catalunya 2004.

[23] Trivedi P, Zimmer D: **Copula modeling: An introduction for practitioners**. *Foundation and trends in Econometrics* 2007, **1**:1–111.

[24] Porta N: **Interval-censored semi-competing risks data: A novel approach for modeling bladder cancer** Tesi doctoral Universitat Politècnica de Catalunya 2010.

[25] Beyersmann J, Latouche A, Buchholz A, Schumacher M: **Simulating competing risks data in survival analysis**. *Stat Med* 2009, **28**:956–971.

[26] Bush D, Smith B, Younger J, Michaelson JS: **The non-breast-cancer death rate among breast cancer patients**. *Breast Cancer Res Treat* 2010.

[27] Gomez G, Delicado P: **Curso de inferencia y decisión**. Tech. rep., Universitat Politècnica de Catalunya 2006.

[28] Curado M, Edwards B, Shin H, Storm H, Ferlay J, Heanue M: **Cancer incidence in five continents, IX**. Tech. Rep. 160, IARC Scientific Publications 2007.

[29] NCI S: **Lifetime risk**. *http://seer.cancer.gov/, accedit gener de 2011.*

[30] Bunker JP, Houghton J, Baum M: **Putting the risk of breast cancer in perspective**. *BMJ* 1998, **317**:1307–1309.

[31] CancerStats: *http://info.cancerresearchuk.org/cancerstats/types/breast/, accedit gener de 2011.*

[32] Reeves GK, Beral V, Bull D, Quinn M: **Estimating relative survival among people registered with cancer in England and Wales**. *Br J Cancer* 1999, **79**:18–22.

Apèndix A

1. Codi de la Simulació

```
library(survival)
source("leadtime.r")
source("sens_mst.r")
source("fun_incidencia.r")
source("fun_mortalitat.r")
source("fun_supervivencia.r")
source("search.lambda.r")

###Dades d'incidència i mortalitat
N<-100000 #mida de la cohort
cohort<-1950 #any naixement cohort

incidence<-function(x)EIB(x,cohort,coef) #incidència CM background
delta<-1
tis<-seq(0,120,delta) #edats i intervals d'edat
nt<-length(tis)
A<-cbind(tis[-nt],tis[-1])

I<-numeric() #incidencia CM agrupada per grups d'edat
for(i in 1:nrow(A))I[i]<-integrate(incidence,A[i,1],A[i,2])$value

tm<-numeric() #risc de morir per altres causes agrupat per grups d'edat
for(i in 1:nrow(A))tm[i]<-integrate(hAC,A[i,1],A[i,2])$value


for (B in 1:500){
NI<-numeric() #nº persones incidents durant l'interval d'edat
NM<-numeric() #nº persones mortes per altres causes durant l'interval d'edat
NVf<-numeric() #nº persones vives al final de l'interval

NI[1]<-rpois(1,N*I[1])
NM[1]<-rpois(1,N*tm[1])
NVf[1]<-N-NI[1]-NM[1]

i<-2
while(NVf[i-1]>0){
NI[i]<-rpois(1,NVf[i-1]*I[i])
NM[i]<-rpois(1,NVf[i-1]*tm[i])
```

```
if(NVf[i-1]-NI[i]-NM[i]<0){
NVf[i]<-0
NM[i]<-NVf[i-1]-NI[i]
}
else{
NVf[i]<-NVf[i-1]-NI[i]-NM[i]
}
i<-i+1
}

NVi<-c(N,NVf[-length(NVf)]) #nº persones vives inici de l'interval
maxedat<-i-1 #extrem superior de l'ultim interval d'edat on hi ha morts
n<-sum(NI) #calcul dels nombre total d'incidents

x0<-numeric() #edats d'incidència
if (n>0){
for(j in 1:length(NI)){
aux<-runif(NI[j],A[j,1],A[j,2])
x0<-c(x0,aux)
}
if(length(x0)>1)x0<-sample(x0)
}

####Simulació temps de supervivència####

##Simulacio 25-39
x025<-x0[x0<40]
n25<-length(x025)
cens25<-numeric() #indicador censura
X25<-numeric() #temps de supervivencia clinic
prob<-numeric() #probabilitat de la causa cancer de mama
for(i in 1:n25){
r<-runif(1)
f<-function(x,r)ST25(x,x025[i])-r #distribució truncada per cada observació
X25[i]<-uniroot(f,c(0,100*(1-r)),r=r)$root #transformada inversa
prob[i]<-(h25(X25[i]))/(h25(X25[i])+hAC(X25[i]+x025[i])) #probabilitat causes
cens25[i]<-rbinom(1,1,prob[i]) #assignem les causes i censurem
}

##Simulacio 40-49
x040<-x0[(x0>=40)&(x0<50)]
n40<-length(x040)
cens40<-numeric() #indicador censura
X40<-numeric() #temps de supervivencia clinic
prob<-numeric() #probabilitat de la causa cancer de mama
for(i in 1:n40){
r<-runif(1)
f<-function(x,r)ST40(x,x040[i])-r #distribució truncada per cada observació
X40[i]<-uniroot(f,c(0,(1-r)*80),r=r)$root #transformada inversa
prob[i]<-(h40(X40[i]))/(h40(X40[i])+hAC(X40[i]+x040[i])) #probabilitat causes
cens40[i]<-rbinom(1,1,prob[i]) #assignem les causes i censurem
}

##Simulacio 50-59
x050<-x0[(x0>=50)&(x0<60)]
n50<-length(x050)
cens50<-numeric() #indicador censura
X50<-numeric() #temps de supervivencia clinic
```

```
prob<-numeric() #probabilitat de la causa cancer de mama
for(i in 1:n50){
r<-runif(1)
f<-function(x,r)ST50(x,x050[i])-r #distribució truncada per cada observació
X50[i]<-uniroot(f,c(0,70*(1-r)),r=r)$root #transformada inversa
prob[i]<-(h50(X50[i]))/(h50(X50[i])+hAC(X50[i]+x050[i])) #probabilitat causes
cens50[i]<-rbinom(1,1,prob[i]) #assignem les causes i censurem
}

##Simulacio 60-69
x060<-x0[(x0>=60)&(x0<70)]
n60<-length(x060)
cens60<-numeric() #indicador censura
X60<-numeric() #temps de supervivencia clinic
prob<-numeric() #probabilitat de la causa cancer de mama
for(i in 1:n60){
r<-runif(1)
f<-function(x,r)ST60(x,x060[i])-r #distribució truncada per cada observació
X60[i]<-uniroot(f,c(0,60*(1-r)),r=r)$root #transformada inversa
prob[i]<-(h60(X60[i]))/(h60(X60[i])+hAC(X60[i]+x060[i])) #probabilitat causes
cens60[i]<-rbinom(1,1,prob[i]) #assignem les causes i censurem
}

##Simulacio 70-85
x070<-x0[x0>=70]
n70<-length(x070)
cens70<-numeric() #indicador censura
X70<-numeric() #temps de supervivencia clinic
prob<-numeric() #probabilitat de la causa cancer de mama
for(i in 1:n70){
r<-runif(1)
f<-function(x,r)ST70(x,x070[i])-r #distribució truncada per cada observació
X70[i]<-uniroot(f,c(0,50*(1-r)),r=r)$root #transformada inversa
prob[i]<-(h70(X70[i]))/(h70(X70[i])+hAC(X70[i]+x070[i])) #probabilitat causes
cens70[i]<-rbinom(1,1,prob[i]) #assignem les causes i censurem
}

ncm<-sum(cens25)+sum(cens40)+sum(cens50)+sum(cens60)+sum(cens70)

####Analisi supervivencia####

##Supervivencia grups d'edat
##N_A
X0<-c(x025,x040,x050,x060,x070)
X<-c(X25,X40,X50,X60,X70)
C<-c(cens25,cens40,cens50,cens60,cens70)
status<-rep(1,n)
E<-cut(X0,c(0,40,50,60,70,120))
svfnae2<-survfit(Surv(X,status)~E,type='fh2')

##Ajustem una distribució per a les diferents supervivencies
##Grup 25-39
SC25<-splinefun(c(0,svfnae2[1]$time),c(1,svfnae2[1]$surv),method="monoH.FC")
SCI25<-function(x,r)SC25(x)-r
##Grup 40-49
SC40<-splinefun(c(0,svfnae2[2]$time),c(1,svfnae2[2]$surv),method="monoH.FC")
SCI40<-function(x,r)SC40(x)-r
##Grup 50-59
```

```
SC50<-splinefun(c(0,svfnae2[3]$time),c(1,svfnae2[3]$surv),method="monoH.FC")
SCI50<-function(x,r)SC50(x)-r
##Grup 60-69
SC60<-splinefun(c(0,svfnae2[4]$time),c(1,svfnae2[4]$surv),method="monoH.FC")
SCI60<-function(x,r)SC60(x)-r
##Grup 70-85
SC70<-splinefun(c(0,svfnae2[5]$time),c(1,svfnae2[5]$surv),method="monoH.FC")
SCI70<-function(x,r)SC70(x)-r

####COPULES####
##Grup 25-39
alp<-3
lambda<-1/sapply(x025,mst)
TC25<-numeric()
TS25<-numeric()
v1<-SC25(X25)
v2<-runif(n25)
u1<-v1
u2<-((v1^(1-alp))*((v2^((1-alp)/alp))-1)+1)^(1/(1-alp))
for(i in 1:n25){TC25[i]<-uniroot(SCI25,c(0,120),r=u1[i])$root}
TS25<-(-1/lambda)*log(u2)
#TS25<-search.lambda(n25, x025, TS25, u2)
##Grup 40-49
alp<-3
lambda<-1/sapply(x040,mst)
TC40<-numeric()
TS40<-numeric()
v1<-SC40(X40)
v2<-runif(n40)
u1<-v1
u2<-((v1^(1-alp))*((v2^((1-alp)/alp))-1)+1)^(1/(1-alp))
for(i in 1:n40){TC40[i]<-uniroot(SCI40,c(0,100),r=u1[i])$root}
TS40<-(-1/lambda)*log(u2)
#TS40<-search.lambda(n40, x040, TS40, u2)
##Grup 50-59
alp<-3
lambda<-1/sapply(x050,mst)
TC50<-numeric()
TS50<-numeric()
v1<-SC50(X50)
v2<-runif(n50)
u1<-v1
u2<-((v1^(1-alp))*((v2^((1-alp)/alp))-1)+1)^(1/(1-alp))
for(i in 1:n50){TC50[i]<-uniroot(SCI50,c(0,100),r=u1[i])$root}
TS50<-(-1/lambda)*log(u2)
#TS50<-search.lambda(n50, x050, TS50, u2)
##Grup 60-69
alp<-3
lambda<-1/sapply(x060,mst)
TC60<-numeric()
TS60<-numeric()
v1<-SC60(X60)
v2<-runif(n60)
u1<-v1
u2<-((v1^(1-alp))*((v2^((1-alp)/alp))-1)+1)^(1/(1-alp))
for(i in 1:n60){TC60[i]<-uniroot(SCI60,c(0,100),r=u1[i])$root}
TS60<-(-1/lambda)*log(u2)
#TS60<-search.lambda(n60, x060, TS60, u2)
```

```
##Grup 70-85
alp<-1.5
lambda<-1/sapply(x070,mst)
TC70<-numeric()
TS70<-numeric()
v1<-SC70(X70)
v2<-runif(n70)
u1<-v1
u2<-((v1^(1-alp))*((v2^((1-alp)/alp))-1)+1)^(1/(1-alp))
for(i in 1:n70){TC70[i]<-uniroot(SCI70,c(0,100),r=u1[i])$root}
TS70<-(-1/lambda)*log(u2)
#TS70<-search.lambda(n70, x070, TS70, u2)


### Resum temps
TS<-c(TS25,TS40,TS50,TS60,TS70)
TC<-c(TC25,TC40,TC50,TC60,TC70)
T0<-X0-TS
D<-X0+TC
DCM<-c()
DAC<-c()
DCM[C==1]<-D[C==1]
DAC[C==0]<-D[C==0]

###Programa de cribratge

for(k in 1:4){
ecrib<-c(40,40,50,50) #edat començament cribratge
nex<-c(15,30,10,20)  #n° examens
tex<-c(2,1,2,1)    #temps entre examens (1,2)
crib<-seq(ecrib[k],ecrib[k]+(nex[k]-1)*tex[k],by=tex[k]) #sequencia d'examens
p<-sapply(crib,sensibilitat) #sensibilitat de l'examen (lee2006)
lcrib<-list(ecrib=ecrib[k],nex=nex[k],tex=tex[k],crib=crib,p=p)
###Lead-Time

lead<-leadtime(lcrib,T0,X0)
edat_det<-lead$edet
lead_time<-lead$lt
temps<-data.frame(T0,TS,X0,TC,DCM,DAC,C,edat_det,lead_time)
write.table(temps, file = paste("Ttemps_",k,"_",B,".txt",sep=""))
}
individus<-data.frame(NVi,NI,NM,NVf,row.names = paste(0:(maxedat-1),
1:maxedat, sep = "-"))
write.table(individus, file = paste("Tindividus_",B,".txt",sep=""))
}
```