# Quantifying the natural history of breast cancer

K H X Tan[1], L Simonella[1], H L Wee[2], A Roellin[3], Y-W Lim[1], W-Y Lim[1,4], K S Chia[1], M Hartman[1,5] and A R Cook*,[1,3,6,7]

[1]Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Block MD3, Level 3, 16 Medical Drive, Singapore 117597, Singapore; [2]Department of Pharmacy, National University of Singapore, Block S4, Level 2, 18 Science Drive 4, Faculty of Science, Singapore 117543, Singapore; [3]Department of Statistics and Applied Probability, National University of Singapore, Block S16, Level 7, 6 Science Drive 2, Faculty of Science, Singapore 117546, Singapore; [4]Ministry of Health, Singapore, College of Medicine Building, 16 College Road, Singapore 169854, Singapore; [5]Department of Surgery, National University Hospital, National University Health System, 5 Lower Kent Ridge Road, Singapore 119074, Singapore; [6]Program in Health Services and Systems Research, Duke-NUS Graduate Medical School Singapore, 8 College Road, Singapore 169857, Singapore and [7]Yale-NUS College, National University of Singapore, 6 College Avenue East, Singapore 138614, Singapore

**Background:** Natural history models of breast cancer progression provide an opportunity to evaluate and identify optimal screening scenarios. This paper describes a detailed Markov model characterising breast cancer tumour progression.

**Methods:** Breast cancer is modelled by a 13-state continuous-time Markov model. The model differentiates between indolent and aggressive ductal carcinomas *in situ* tumours, and aggressive tumours of different sizes. We compared such aggressive cancers, that is, which are non-indolent, to those which are non-growing and regressing. Model input parameters and structure were informed by the 1978–1984 Ostergotland county breast screening randomised controlled trial. Overlaid on the natural history model is the effect of screening on diagnosis. Parameters were estimated using Bayesian methods. Markov chain Monte Carlo integration was used to sample the resulting posterior distribution.

**Results:** The breast cancer incidence rate in the Ostergotland population was 21 (95% CI: 17–25) per 10 000 woman-years. Accounting for length-biased sampling, an estimated 91% (95% CI: 85–97%) of breast cancers were aggressive. Larger tumours, 21–50 mm, had an average sojourn of 6 years (95% CI: 3–16 years), whereas aggressive ductal carcinomas *in situ* took around half a month (95% CI: 0–1 month) to progress to the invasive ≤10 mm state.

**Conclusion:** These tumour progression rate estimates may facilitate future work analysing cost-effectiveness and quality-adjusted life years for various screening strategies.

Decision analytic models have become a cornerstone in assessing costs and benefits of pharmaceutical and health technology interventions (Brennan and Akehurst, 2000), allowing the benefits of an intervention evaluated in clinical trials to be estimated relative to standard care, as well as to determine the implications for modification of an intervention's use (Sun and Faunce, 2008). For breast cancer, a number of randomised controlled trials have

highlighted that biennial mammographic screening reduces mortality due to breast cancer among women aged 50–69 years (Independent UK Panel on Breast Cancer Screening, 2012). However, the number of women who need to be screened to prevent one death has become an issue of debate (Independent UK Panel on Breast Cancer Screening, 2012). Detailed natural history models of breast cancer progression provide an opportunity to

BJC OPEN

evaluate modification to screening programmes to identify optimal screening scenarios while minimising screening's negative effects, such as unnecessary biopsies following false positive tests and treatment of indolent cancers (Kobrunner *et al*, 2011).

A number of breast cancer natural history models have been developed. Three-state Markov models developed by Tabar *et al* (1995), Duffy *et al* (1995), Duffy *et al* (1997) and Wu *et al* (2010) permit simulation of the natural history of breast cancer by characterising the disease process, starting from non-diseased to preclinical cancer, then clinical cancer states. More complex models evolved from these, including a five-state Markov model, which differentiates localised from non-localised tumours (Wu *et al*, 2010) and a model by Duffy *et al* (1997) which includes nodal involvement for preclinical and clinical disease. However, although these models provide a broad overview of the natural history of breast cancer, the heterogeneous nature of breast cancer, in terms of treatment and prognosis, at different stages of its growth may lead to these not being sufficient for detailed health economic evaluation of the impact of modified breast screening scenarios.

Breast cancer size is highly predictive of survival (Narod, 2012) and strongly influences the mode of treatment (Kurian *et al*, 2012), so health economic models that invoke earlier detection at smaller sizes combined with survival analysis of outcomes need reliable estimates of the growth rates of tumours of varying sizes. A detailed understanding of tumour progression could serve as a starting point to evaluate refined screening algorithms (Duffy *et al*, 1995; Plevritis *et al*, 2007). Our aims in this paper are to develop a detailed model characterising breast cancer tumour progression utilising a randomised controlled trial in a population without pre-existing exposure to breast screening and to evaluate the effects on the distribution of tumour sizes at detection of screening strategies varying in screening frequency and breast cancer incidence.

## MATERIALS AND METHODS

**Data sources.** In 1978, all women aged 40 years or more in the county of Ostergotland, Sweden, were randomised to either invitation to participate in screening or to what was then the standard care (no screening) (Fagerberg *et al*, 1985). Figure 1 (and Supplementary Material Table 7) outlines the observed screening pathway and detected tumour sizes.

In the screening arm, 38 496 women were invited to participate, excluding those older than 74, who had participation rates that were not as complete as younger women. The time from randomisation to first screen was $\sim 1$ month, whereas the average interval between the first and second screen was 27 months. The model categorised women in the screening arm into one of 10 outcome scenarios (Figure 1A) accounting for: detection (by the patient or a clinician), detection at screening, attendance at the two screens, and the time frame of detection (before screen 1, between the screens, or at either screen). Diagnosed tumours are categorised by size: ductal carcinomas *in situ* (DCIS), $\leqslant 10$ mm, 11–20 mm, 21–50 mm, and $\geqslant 51$ mm. The data do not, however, differentiate results at screen 2 by attendance at screen 1, necessitating data augmentation techniques (see Supplementary Material for details). The model used (see next section) implicitly defines the probability of each tumour size in each of these 10 scenarios. In the control arm, there were 37 936 women (under 75) for whom there were two outcomes: asymptomatic or symptomatic disease (by size) by the end of the study period.

Data on screening sensitivity, that is, the probability of positive test results in a woman with an undiagnosed tumour, by tumour size, were taken from a study in northern California (Kerlikowske *et al*, 1996). In addition, we used the proportion of detected DCIS

that were invasive in a review of eight studies (Leonard and Swain, 2004).

**Breast cancer natural history model structure.** We constructed a 13-state continuous-time Markov model with 10 transition parameters (Figure 1C; a full list of parameters may be found in Supplementary Materials Table 4). The model differentiates between indolent and aggressive DCIS tumours, and aggressive tumours of different sizes. Here we use the term indolent to mean a DCIS that never progresses, and aggressive to mean a DCIS that will progress. The transition states begin from 'no cancer' to 'indolent DCIS' or 'aggressive DCIS', with subsequent progression from 'aggressive DCIS' to larger tumour sizes. Rates of progression for each stage of disease depend on the current size of the tumour (DCIS, $\leqslant 10$, 11–20, 21–50, and $\geqslant 51$ mm), and the corresponding four parameters, which define mean sojourn times in each size class, permit a range of behaviours, including some that mimic exponential or Gompertzian growth. In the absence of mammography, symptomatic disease is detected at certain stages of disease at rates that also depend on current tumour size. At the point of detection, the woman enters an absorbing diseased state.

The transition parameters and model framework allow 'absorption probabilities' of tumours at different sizes (i.e. the proportion of individuals who would end in one of the symptomatic diagnosed states, given an infinite period of time) to be characterised (Figure 1C), along with the steady-state proportions of women with undiagnosed cancer, in which we condition on no diagnosis over an infinite time horizon, which corresponds to the equilibrium distribution of cancer presence, and size if present, among asymptomatic women. These proportions allow the long-run distribution of tumour sizes in asymptomatic and symptomatic women to be derived from the individual components of the model.

**Mammography model structure.** Overlaid on the natural history model is the effect of screening on diagnosis. Women without symptomatic disease have a specific probability of attending screening, assumed to be the same for those with no or asymptomatic disease. Those with cancer are diagnosed with a probability that depends on the size of the tumour, and are removed from the natural history model if cancer is diagnosed, whereas those with no detected cancer continue through the model until symptomatic detection or detection at the second screen. Test characteristics for screening by tumour size were derived from Kerlikowske *et al* (1996).

**Parameter estimation.** Parameter estimates for the model were obtained directly, and primarily, from the Ostergotland trial using Bayesian methods, with other data entering the analysis via an informative prior distribution. We used Markov chain Monte Carlo integration (Albert, 2007) to sample the resulting, non-analytically tractable, posterior distribution for the 36 parameters, (listed in Table 2 in Supplementary Material).

**Likelihood function.** The likelihood function was calculated for both the screening and control study arms, as the product of two multinomial probability masses. For each arm, there was one outcome for each scenario and tumour size, the probabilities of which, as functions of the parameters, were derived directly from the Markov model formulation. Specifically, a $13 \times 13$ transition rate matrix $Q(\theta)$ was formed, and the distribution of the model state, $X_t$, at time $t$ years after the start of the study, as a function of the parameters $\theta$ was derived using the relationship $p(X_t|\theta) = p(X_0|\theta)\exp\{Q(\theta)t\}$, where $\exp\{ \}$ is the matrix exponential function (Zwillinger, 2011). Using this relationship, along with the average time intervals between randomisation to first screen ($T_{01}$), first screen to second ($T_{12}$) and randomisation to second screen ($T_{02}$), the probabilities of being in different scenarios, and tumour
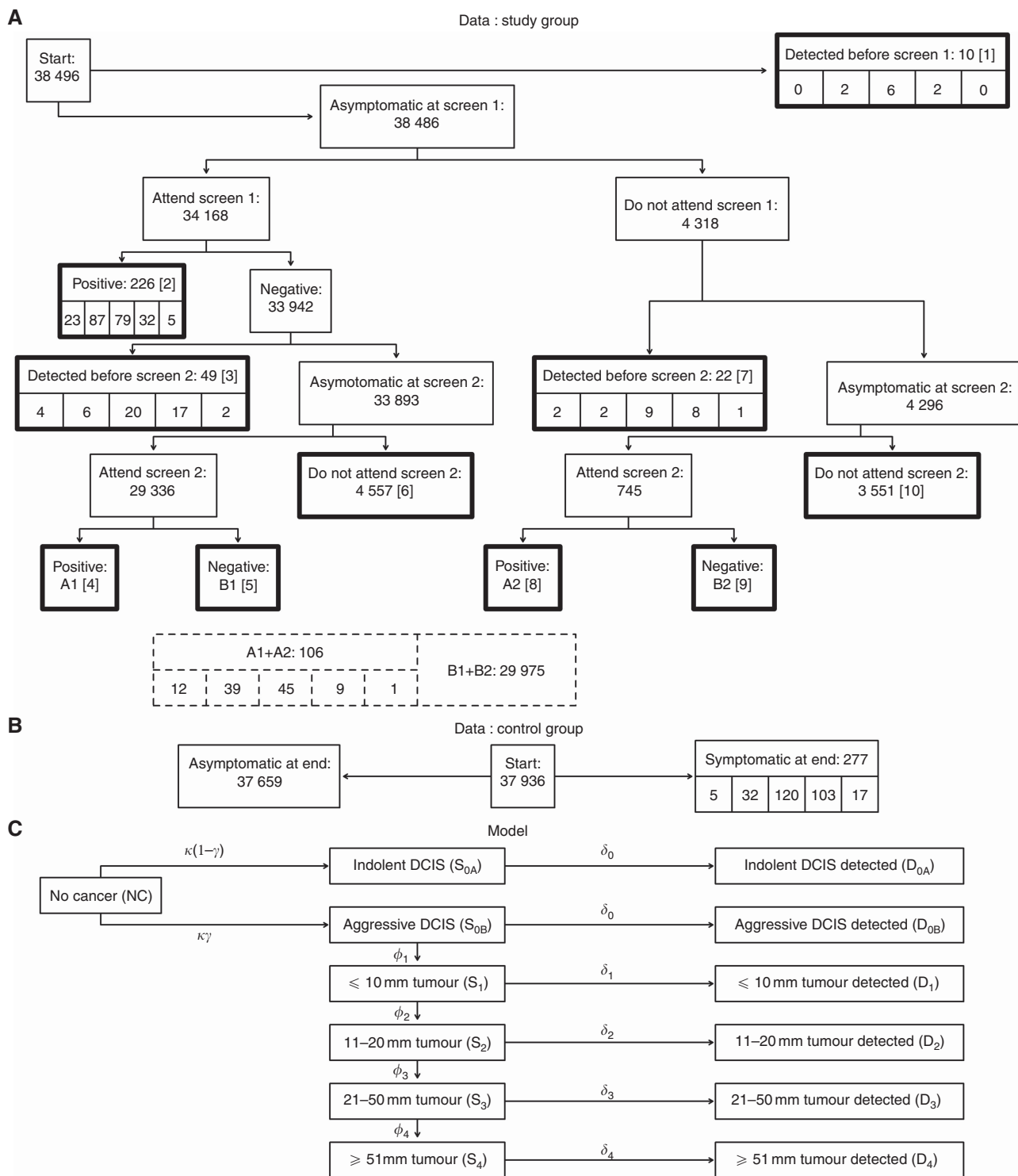
**Figure 1. Flow diagrams of the data used in the study and 13-state Markov model in the absence of screening.** (**A**) Ten outcomes are marked in thickly bordered rectangles with tumour sizes, in order DCIS, ⩽10, 11–20, 21–50 and ⩾51 mm, underneath. Data were extracted from tables and text of Fagerberg *et al* (1985). The publication did not distinguish screening results at the second screen by attendance at the first screen: the merged data are provided in the dashed rectangles at the foot. (**B**) Asymptomatic and symptomatic controls at end of the study, with data for each tumour size group given underneath. (**C**) Greek letters indicate transition rates from one state to another. The detection states are absorbing: once a woman enters those, she goes for treatment and does not progress within this model.

sizes, in screened and control groups were calculated as a deterministic function of the unknown parameters.

For the control arm, the probabilities of symptomatic disease by size, or no symptomatic disease, were extracted from relevant entries of $p(X_{T_{02}} \mid \theta)$. For the screening arm, the derivation was more complicated and involved conditioning the probability of subsequent states upon previous ones and integrating over unknown past states. In all, including the different size distributions, there are 43 equations to derive the likelihood function, detailed in full in Supplementary Material.

**Prior distributions for parameters.** Informative prior distributions for tumour-dependent screening sensitivity (Table 6 in Supplementary Material) were developed by fitting a beta distribution to the sample size and number of cases in each tumour size stratum (Table 3 in Supplementary Material), exploiting conjugacy between the beta and binomial distributions (Gelman *et al*, 2004). We incorporated external information on the prevalence of indolent DCIS using Bayesian melding (Poole and Raftery, 2000) by setting the prior distribution for the probability of getting aggressive breast cancer to be non-informative (i.e. $U(0,1)$) and incorporating an additional term in the posterior for each screen in which DCIS could be detected (see Supplementary Material), with a parameter characterising the prevalence of aggressive DCIS on screening, with an informative prior distribution derived from Leonard and Swain (2004). Transition rates – that is, for incidence, growth, symptomatic disease – had Exp(0.01) prior distributions to ensure identifiability, whereas all other parameters – that is, for attendance at screening and initial proportions in each state – had uniform prior distributions over the parameter support (see Supplementary Material).

**Posterior distributions for parameters and initial conditions.** The posterior distribution for the model parameters was sampled using the Metropolis–Hastings algorithm (Metropolis *et al*, 1953; Hastings, 1970), with univariate proposal distributions for all parameters except the initial states, which used a multivariate normal proposal distribution. Tuning parameters were selected on pilot studies. Four independent chains, of 52 000 iterations each, were run in parallel, with 2000 iterations dropped as burn-in. Point estimates are posterior means, and intervals are equal-tailed credible intervals, unless otherwise noted. The 95% credible interval for each parameter, shown in Table 1, was determined by taking the 2.5th and 97.5th centiles from the posterior sample. Convergence of the Markov chain Monte Carlo samplers was assessed graphically and using the Gelman–Rubin diagnostic (Gelman and Rubin, 1992) in the CODA package (Plummer *et al*, 2006). The software used was R Version 2.15.2 (Venables and Ripley, 2002; Goulet *et al*, 2012; R Core Team, 2012); an R script to fit the 13-state model is provided in the Supplementary Material.

**Varying mammographic screening frequency and breast cancer risk.** We assessed how the posterior predictive distribution of tumour sizes changed with the frequency of mammographic screening over a 10-year time horizon, from no screening to annual, biennial or quinquennial. Holding screening fixed at a biennial frequency, we also varied the underlying risk of breast cancer, by scaling the breast cancer incidence rate, and derived the tumour size distribution. Low (50% of baseline), normal (100%), moderate (150%) and high (200%) risk were considered. Different risk levels were explored to inform follow-on studies that extrapolate to other populations with different incidence rates or assess tailored screening programmes for different risk groups.

**Sensitivity analysis.** A sensitivity analysis was performed using an 11-state Markov model, which does not differentiate indolent DCIS from aggressive. In this smaller model, all DCIS cases are assumed able to progress to invasive, with all other states remaining the same (Figure 1C). This was done to assess the sensitivity of our findings to the assumption that some DCIS are indolent and never progress to aggressive tumours.

## RESULTS

**Model validation.** Internal validation of the 13-state Markov model was done by plotting the posterior predictive (i.e. modelled) distribution of proportions against the data with their 95% classical Wald confidence intervals (Figure 2), indicating the close fit.

**Parameter estimates.** Based on the 13-state Markov model (Table 1a), the breast cancer incidence rate in the women of Ostergotland aged 40–74 was 21 (95% CI: 17–25) per 10 000 woman-years, with 2 (95% CI: 1–3) and 19 (95% CI: 16–23) per 10 000 woman-years for indolent and aggressive breast cancers, respectively. An estimated 91% (95% CI: 85–97%) of breast cancers were aggressive. Larger tumours had longer sojourn times, with an average sojourn of 6 years (95% CI: 3–16 years) between 21 and 50 mm, whereas aggressive DCIS took an estimated mere half a month (95% CI: 0–1 month) to progress to the invasive ⩽10 mm state. The mean time spent in ⩽10, 11–20, and ⩾51 mm before progression (or detection) was about 10 months (95% CI: 7–13 months), 2 years (95% CI: 2–4 years), and 5 months (95% CI: 1–11 months: n.b. this corresponds to detection only), respectively. Almost no aggressive DCIS were detected before progression (0%; 95% CI: 0–1%) but the probability of detecting ⩽10 mm before it progressed to 11–20 mm was 12% (95% CI: 8–15%), from 11–20 to 21–50 mm was 51% (95% CI: 43–60%), and from 21–50 to ⩾51 mm was 87% (95% CI: 79–95%). The estimated screening sensitivity ranged from 88 to 93% (Table 1a). In the absence of screening, with all breast cancers detected by other means, the proportion detected with a DCIS was 9% (95% CI: 4–15%), ⩽10 mm was 10% (95% CI: 8–14%), 11–20 mm was 41% (95% CI: 35–48%), 21–50 mm was 34% (95% CI: 28–41%), and ⩾51 mm was 5% (95% CI: 2–9%). At any instant of time, of those with asymptomatic cancer, 30% (95% CI: 12–53%) will have indolent DCIS, 2% (95% CI: 0–4%) aggressive DCIS, 24% (95% CI: 15–33%) ⩽10 mm, 33% (95% CI: 21–47%) 11–20 mm, 10% (95% CI: 5–18%) 21–50 mm, and 1% (95% CI: 0–2%) ⩾51 mm. Histograms of the marginal posterior distributions may be found in Supplementary Materials Figure 7.

Different mammographic screening frequencies (Figures 3A–D) and breast cancer risks (Figures 3E–H) were explored independently. Under different screening frequencies, the more significant impacts are observed in the proportions of ⩽10 mm and 21–50 mm tumours. Over a 10-year time horizon, the proportion of women aged 40–74 diagnosed early, that is, with a ⩽10 mm tumour, increases from 0.3% to 0.4%, 0.7%, and 1.0% as the programme moves from no screening, to quinquennial, biennial, and annual screening, respectively. Similarly, the respective proportion of women diagnosed with a 21–50 mm tumour decreases from 0.8 to 0.7, 0.4, and 0.3%. By introducing annual screening, there is a much higher proportion of women diagnosed with ⩽10 mm tumour (from 0.3 to 1.0%) and a substantially lower proportion of women diagnosed with 21–50 mm tumour (from 0.8 to 0.3%), as compared with no screening. Comparing no screening and biennial screening, there is also an increase in proportion of women having ⩽10 mm tumour and decrease in proportion of women having 21–50 mm tumour, but both the increase (from 0.3 to 0.7%) and the decrease (from 0.8 to 0.4%) are lower than that obtained when comparing annual screening to baseline. For quinquennial screening, although the proportion of small tumours (⩽10 mm) rises from 0.3 to 0.4% and the proportion of larger tumours (21–50 mm) falls from 0.8 to 0.7%, the clinically significant difference in the distribution of tumour sizes observed for more frequent screening is not realised. In contrast, varying the underlying level of breast cancer risk (Figures 3E–H) does little to the distribution of tumour sizes discovered.

**Sensitivity analysis.** Internal validation indicated that the 11-state model (Figure 4 in Supplementary Material) also provides a good fit to the data. There are few differences between corresponding parameters in the two models, the main being longer average sojourns in the DCIS class under the model with no indolent DCIS (0.04 years in 13-state model *vs* 0.2 years in 11-state model).

| Table 1. List of parameter and derived parameter estimates from (a) 13-state model and (b) 11-state model | | | | |
|---|---|---|---|---|
| | **(a) 13-state model** | | **(b) 11-state model** | |
| **Parameters** | **Estimates** | **95% CI** | **Estimates** | **95% CI** |
| Probability (%) of getting aggressive breast cancer | 91 | 85–97 | — | — |
| **Incidence rate (per 10 000 woman-years)** | | | | |
| Breast cancer | 21 | 17–25 | 21 | 17–25 |
| Indolent breast cancer | 2 | 1–3 | — | — |
| Aggressive breast cancer | 19 | 16–23 | — | — |
| **Initial probability (%)** | | | | |
| No cancer | 99.23 | 99.13–99.32 | 99.24 | 99.14–99.33 |
| Indolent DCIS | 0.06 | 0.03–0.09 | 0.10 | 0.05–0.15 |
| Aggressive DCIS | 0.13 | 0.01–0.32 | 0.10 | 0.05–0.15 |
| ≤10 mm | 0.21 | 0.02–0.35 | 0.29 | 0.22–0.37 |
| 11–20 mm | 0.26 | 0.20–0.33 | 0.25 | 0.20–0.32 |
| 21–50 mm | 0.10 | 0.07–0.14 | 0.10 | 0.10–0.14 |
| ≥51 mm | 0.02 | 0.00–0.03 | 0.02 | 0.00–0.03 |
| **Average sojourn time, in years, for different tumour sizes** | | | | |
| Aggressive DCIS | 0.0 | 0.0–0.1 | 0.2 | 0.1–0.4 |
| ≤10 mm | 0.8 | 0.6–1.1 | 0.8 | 0.6–1.1 |
| 11–20 mm | 2.4 | 1.6–3.5 | 2.4 | 1.6–3.4 |
| 21–50 mm | 6.4 | 2.5–15.6 | 6.4 | 2.6–15.5 |
| ≥51 mm | 0.4 | 0.1–0.9 | 0.5 | 0.1–0.9 |
| **Probability (%) of detecting breast cancer before progression in tumour size** | | | | |
| Aggressive DCIS → ≤10 mm | 0 | 0–1 | 3 | 1–4 |
| ≤10 mm → 11–20 mm | 12 | 8–15 | 11 | 8–15 |
| 11–20 mm → 21–50 mm | 51 | 43–60 | 51 | 43–59 |
| 21–50 mm → ≥51 mm | 87 | 79–95 | 87 | 78–95 |
| **Probability (%) of attending screenings** | | | | |
| First screening | 89 | 88–89 | 89 | 88–89 |
| Both screenings | 87 | 86–87 | 87 | 86–87 |
| Second screening but not first screening | 17 | 16–19 | 17 | 16–19 |
| **Sensitivity (%) of mammography for different tumour sizes** | | | | |
| DCIS | 88 | 83–92 | 88 | 83–92 |
| ≤10 mm | 90 | 86–93 | 90 | 87–93 |
| 11–20 mm | 91 | 88–94 | 91 | 88–94 |
| 21–50 mm | 92 | 89–95 | 92 | 89–95 |
| ≥51 mm | 93 | 90–96 | 93 | 90–96 |
| **Absorption probability of different tumour sizes** | | | | |
| DCIS (indolent or aggressive) | 9 | 4–15 | 3 | 1–4 |
| ≤10 mm | 10 | 8–14 | 11 | 8–14 |
| 11–20 mm | 41 | 35–48 | 44 | 37–51 |
| 21–50 mm | 34 | 28–41 | 37 | 30–44 |
| ≥51 mm | 5 | 2–9 | 6 | 2–10 |
| **Steady-state proportion of women with undiagnosed cancer** | | | | |
| No cancer | 99.4 | 99.1–99.6 | 99.5 | 99.4–99.6 |
| Indolent DCIS | 0.2 | 0.1–0.4 | 0.0 | 0.0–0.1 |
| Aggressive DCIS | 0.0 | 0.0–0.0 | 0.0 | 0.0–0.1 |
| ≤10 mm | 0.1 | 0.1–0.2 | 0.1 | 0.1–0.2 |
| 11–20 mm | 0.2 | 0.1–0.3 | 0.2 | 0.1–0.3 |
| 21–50 mm | 0.1 | 0.0–0.1 | 0.1 | 0.0–0.1 |
| ≥51 mm | 0.0 | 0.0–0.0 | 0.0 | 0.0–0.0 |
| **Steady-state proportion of women with undiagnosed cancer (conditional on some cancer)** | | | | |
| Indolent DCIS | 30 | 12–53 | 10 | 6–16 |
| Aggressive DCIS | 2 | 0–4 | 10 | 6–16 |
| ≤10 mm | 24 | 15–33 | 30 | 22–39 |
| 11–20 mm | 33 | 21–47 | 44 | 34–56 |
| 21–50 mm | 10 | 5–18 | 14 | 7–23 |
| ≥51 mm | 1 | 0–2 | 1 | 0–3 |
| Abbreviations: CI = confidence interval; DCIS = ductal carcinomas *in situ*. | | | | |
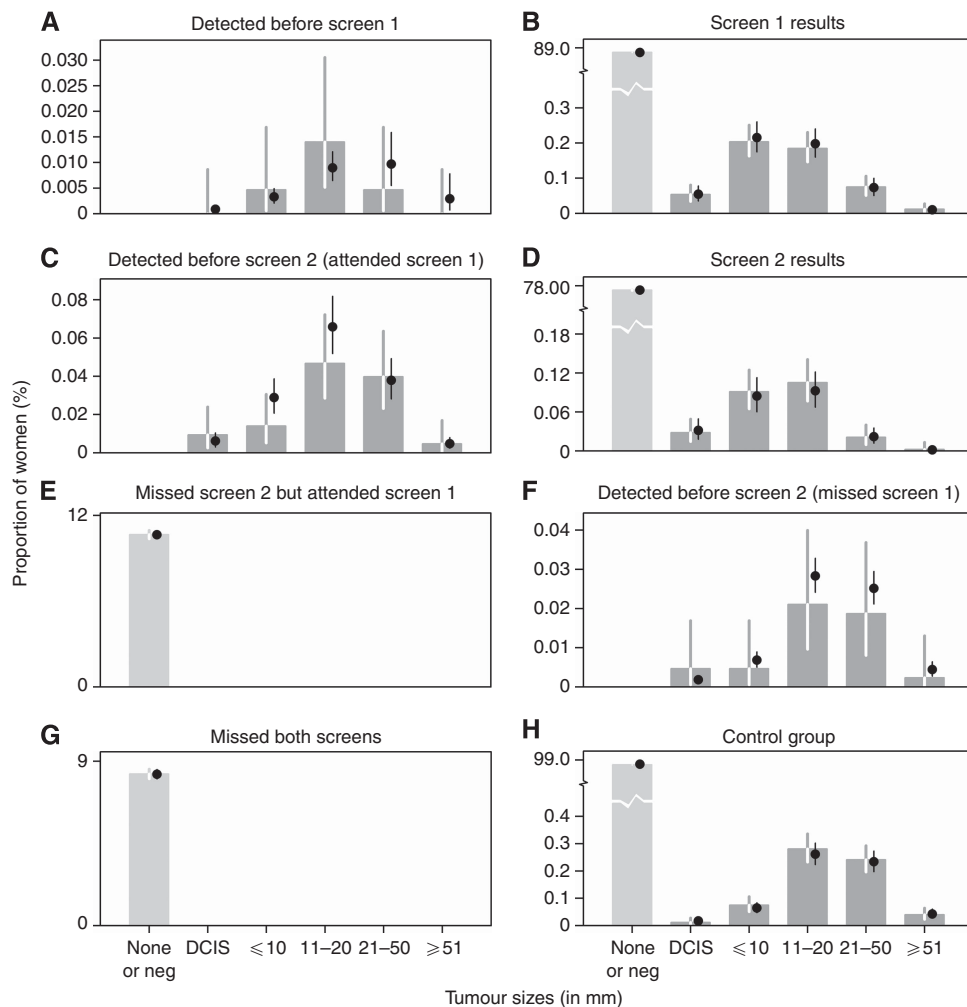
**Figure 2. Data vs predictive distribution of tumour sizes, based on 13-state model.** Bars with lines represent data with their 95% classical confidence intervals based on binomial errors. Points with lines represent modelled proportions and their 95% credible intervals. A close fit between the data and posterior predictive distribution of proportions can be observed, for the various outcomes observed in the data structure (Figure 1**A** and **B**).

## DISCUSSION

(Additional discussion can be found in the Supplementary Material.)

The aim of this analysis was to present a detailed natural history of breast cancer. The data from the Ostergotland study have several characteristics that make it an invaluable resource to understand breast cancer tumour progression. First, at the time of the study, mammography for detection of asymptomatic breast cancer had not yet become established and so the results of the study reflect a relatively 'clean' asymptomatic population without interference from past screens. Second, because it was such an early study, the non-programmatic uptake of mammographic screening during the trial was low (7.5%) in the control arm, which, to a good approximation, can therefore be treated as unscreened. Third, a high proportion of women invited to screening took up the offer (89%), in contrast to the difficulties seen in other populations in getting women to go for routine screening (Fagerberg et al, 1985). Finally, the study design had two rounds of screening which allow the prevalence of existing cancers to be determined in the first round, whereas the second round provided information on the incidence of newly developed tumours.

Previous papers have focused on models of other characteristics such as age (Tabar et al, 1995; Duffy et al, 1997; Straatman et al,

1997) and node status (Duffy et al, 1997; Chen et al, 1998), rather than tumour sizes (see Supplementary Material Table 5). Our modelled breast cancer incidence rate (21 per 10 000 woman-years) is higher (by around 5 per 10 000 woman-years) than the empirical breast cancer incidence in Swedish woman, aged 40–79, during the time period 1978–1984 (Haukka et al, 2011), but comparable to estimates from less-complex models applied to Fennoscandian data (Duffy et al, 1997; Wu et al, 2010). One argument that has been put forward to explain higher long-term breast cancer incidence in screened populations is the potential spontaneous regression of tumours (Kopans et al, 2011). Another is the difference in definitions: incidence in our analysis related to 'cryptic' incidence of unobserved tumours, which the empirical incidence of diagnosed tumours will lag behind.

Some insights provided by our model are ostensibly counter-intuitive. Average progression rate of DCIS to aggressive cancer has been reviewed (Leonard and Swain, 2004) and found to be ~43%, whereas in our model, we estimated a 91% chance of breast cancers to be aggressive – using the same dataset as a prior distribution. The apparent discrepancy is due to our accounting for length-biased sampling (Blumenthal, 1967). This is important, and necessary, as an indolent DCIS, which never progresses, has a substantially greater chance of being detected than a DCIS that is aggressive and has only a short time window to be identified as a DCIS before it grows to an invasive tumour. In fact, to determine
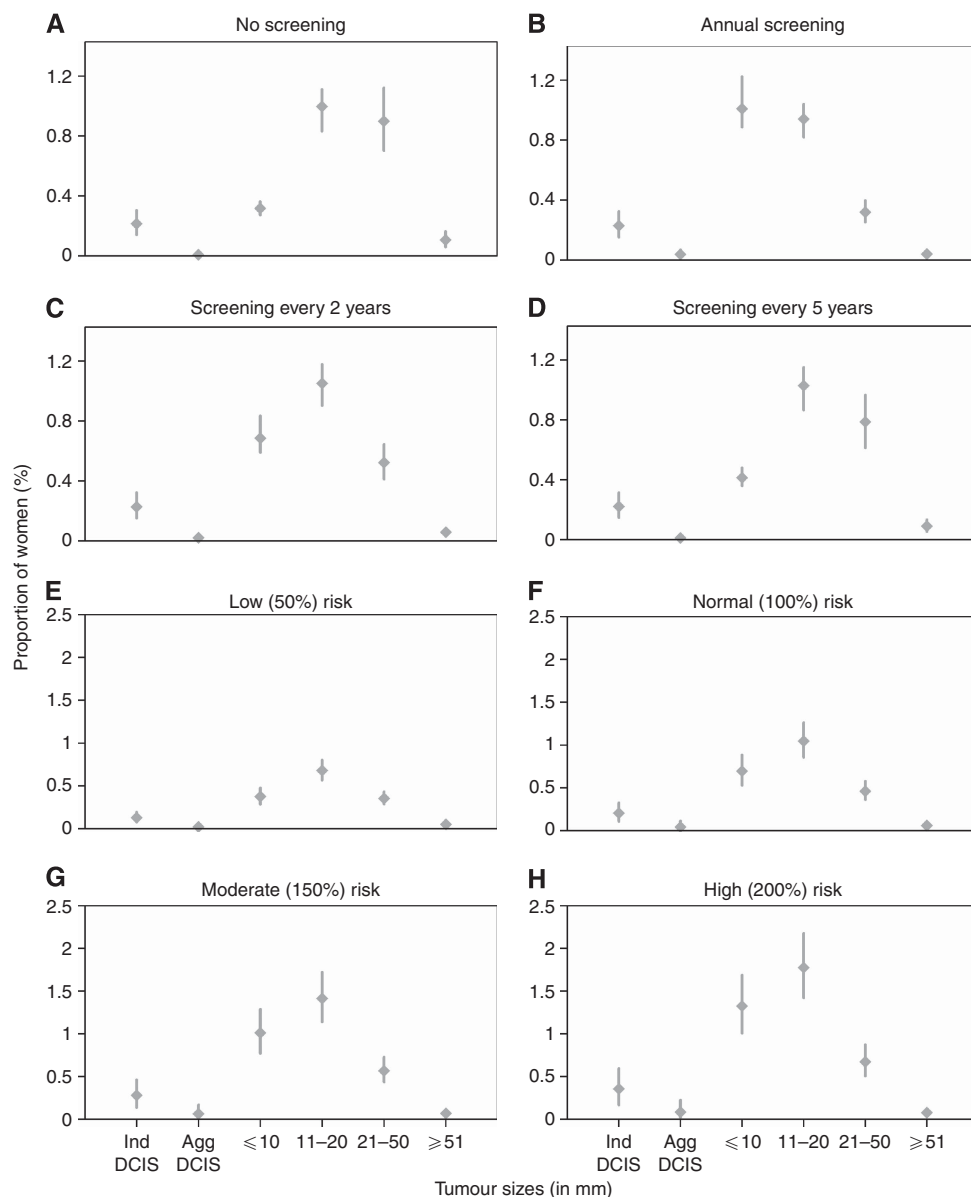
**Figure 3. Tumour size distribution for different mammographic screening frequencies and different rates from no cancer to DCIS, based on 13-state model.** Different mammographic screening frequencies – (**A**) no screening, (**B**) annual screening, (**C**) screening every 2 years, and (**D**) screening every 5 years. Different rates from no cancer to DCIS – (**E**) low risk, (**F**) normal risk, (**G**) moderate risk, and (**H**) high risk.

the true proportion of DCIS that are indolent requires an approach that, like ours, accounts for the duration in which the tumour can be detected as a DCIS. There are few differences in parameter estimates between our 13 and 11-state (i.e. excluding indolent DCIS, see Supplementary Material) models, and both exhibit close concordance to the data, therefore our analysis does not provide evidence to support either hypothesis preferentially, although other studies support the existence of indolent DCIS (Leonard and Swain, 2004). If no DCIS are indolent, then to fit the data, the typical duration with a DCIS before progression to invasive cancer is estimated to be very brief (~2 months). On the other hand, if some DCIS are in fact indolent, never to progress, then those that do progress would have to do so almost instantly to agree with the data from this trial. This has implications for 'over diagnosis', that is, the criticism of mammography that it increases the proportion of detected DCIS that never would progress (Kerlikowske, 2010). Although both models we considered describe the data almost equally well, the simpler one, which excludes non-progressive cancers, necessarily cannot characterise the over diagnosis of

non-progressive cancers, which is a major limitation and one of the motivations for introducing an indolent class of tumours in the main model.

Monte Carlo simulation of the fitted model suggests that screening once every 5 years is not appreciably better than not screening at all, with no clinically significant differences in the tumour size distribution (Figures 3A–D). However, annual and biennial screening do lead to a marked reduction in sizes: over a 10-year period, for 100 women with cancer, annual screening would pick up 16 at a smaller size class, and biennial screening 8, but these would involve an additional eight and three screens, respectively, compared with quinquennial screening. It is therefore important that the cost-effectiveness of these alternative screening frequencies be evaluated in future studies.

An important limitation of our model is that we provide average estimates over all ages in the range 40–74, which was necessary as the original publication does not differentiate outcomes by age. It is, however, necessary to understand the distribution of sojourn times in the targeted age group when formulating screening policy

(Duffy *et al*, 1995), and so future work accounting for different age groups, analysing the cost-effectiveness and looking into women's quality-adjusted life years for various screening strategies, would be valuable. Similarly, mortality due to other causes is not distinguished in the data, and hence in our model, from non-attendance at screens (for the screened arm) or being asymptomatic over the 2 years of the study (in the control arm). This is not a severe limitation due to the short time horizon and age range considered, but for other studies in which follow-up is longer, mortality would need to be incorporated explicitly.

How might our findings generalise – both to the present and to other, especially non-European, populations? Screening technologies have advanced markedly since this early trial, so the screening sensitivities estimated here are no longer relevant. Attendance rates in the Swedish study were higher than those observed since (Lagerlund *et al*, 2000), and self-examination rates may also vary across cultures and over time, thus both mammography uptake and self-detection rates may not generalise well. Transition rates – because they represent a biological process – are, we believe, relevant to both the present and to other settings, though additional validation would be necessary to confirm whether observed differences in breast cancer between ethnic groups affect this (Leong *et al*, 2010). One limitation is that the correlation between age and outcomes was not provided, so our results are averaged over women in the 40–74 age group, but there are known biological differences between premenopausal and postmenopausal cancers, with a quicker progression of disease in younger women (Buist *et al*, 2004). Additional research is needed to account for differing incidence rates and, potentially, probability of getting aggressive breast cancer (Leonard and Swain, 2004). It is also unclear to what extent detection rates, in between screens, will apply to other cultures, as, for example, in developing countries, women tend to present late, at a more advanced disease state (Hortobagyi *et al*, 2005).

Clinical trials of the effectiveness of mammographic screening programmes in reducing mortality were carried out using older technologies, would have led to surgical and medical interventions with poorer prognosis than at present, and were predominantly among ethnic Europeans in whom incidence rates are higher than, say, ethnic East Asians. They also by necessity considered a single frequency of screening (from annual to triennial), from which extrapolation to other frequencies is challenging. *In silico* experimentation allows the evaluation of the effectiveness and cost-effectiveness of screening programmes in settings in which clinical trials have not yet been performed, in women who differ in underlying risk and in acceptance of screening, and in health systems that differ in treatment options and consequent survival. Modelling also permits the evaluation of tailored screening in which women at higher risk within a population are offered more frequent screening, for although we find the distribution of tumour sizes to be rather invariant to incidence in the screened group (Figures 3E–H), the number of cancers found did vary, with implications for cost-effectiveness. The current model focuses entirely on events prior to diagnosis, and future modelling work to evaluate the effectiveness and cost-effectiveness of screening programmes, and to optimise screening strategies accordingly, should extend this to outcomes after diagnosis.

## ACKNOWLEDGEMENTS

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## AUTHORS CONTRIBUTIONS

KHXT designed the study, wrote the computer code, designed the model, performed statistical analyses, and drafted the manuscript. LS designed the study and drafted the manuscript. HLW and WYL helped interpret results and drafted the manuscript. AR and YWL designed the models and drafted the manuscript. KSC conceived the study and drafted the manuscript. MH conceived the study, helped interpret results and drafted the manuscript. ARC conceived the study, wrote the computer code, designed the model, and drafted the manuscript.

## REFERENCES

Albert J (2007) *Bayesian Computation with R.* Springer: New York, USA.

Blumenthal S (1967) Proportional sampling in life length studies. *Technometrics* **9**: 205–218.

Brennan A, Akehurst R (2000) Modelling in health economic evaluation. What is its place? What is its value? *Pharmacoeconomics* **17**(5): 445–459.

Buist DS, Porter PL, Lehman C, Taplin SH, White E (2004) Factors contributing to mammography failure in women aged 40-49 years. *J Natl Cancer Inst* **96**(19): 1432–1440.

Chen HH, Thurfjell E, Duffy SW, Tabar L (1998) Evaluation by Markov chain models of a non-randomised breast cancer screening programme in women aged under 50 years in Sweden. *J Epidemiol Community Health* **52**: 329–335.

Duffy SW, Chen H, Tabar L, Day NE (1995) Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry to and exit from the preclinical detectable phase. *Stat Med* **14**: 1531–1543.

Duffy SW, Day NE, Tabar L, Chen H, Smith TC (1997) Markov Models of breast tumor progression: some age-specific results. *J Natl Cancer Inst Monogr* **1997**(22): 93–97.

Fagerberg G, Baldetorp L, Grontoft O, Lundstrom B, Manson JC, Nordenskjold B (1985) Effects of repeated mammographic screening on breast cancer stage distribution: results from a randomized study of 92 934 women in a Swedish county. *Acta Radiol Oncol* **24**(6): 465–473.

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) *Bayesian Data Analysis,* 2 edn: Chapman & Hall/CRC.

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Stat Sci* **7**: 457–511.

Goulet V, Dutang C, Maechler M, Firth D, Shapira M, Stadelmann Mexpm-developers@lists.R-forge.R-project.org (2012) expm: Matrix exponential.

Hastings W (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**: 97–109.

Haukka J, Byrnes G, Boniol M, Autier P (2011) Trends in breast cancer mortality in Sweden before and after implementation of mammography screening. *PLoS One* **6**(9): e22422.

Hortobagyi GN, de la Garza Salazar J, Pritchard K, Amadori D, Haidinger R, Hudis CA, Khaled H, Liu MC, Martin M, Namer M, O'Shaughnessy JA, Shen ZZ, Albain KS (2005) The global breast cancer burden: variations in epidemiology and survival. *Clin Breast Cancer* **6**: 391–401.

Independent UK Panel on Breast Cancer Screening (2012) The benefits and harms of breast cancer screening: an independent review. *Lancet* **380**: 1778–1786.

Kerlikowske K (2010) Epidemiology of ductal carcinoma *in situ. J Natl Cancer Inst Monogr* **41**: 139–141.

Kerlikowske K, Grady D, Barclay J, Sickles EA, Ernster V (1996) Effect of age, breast density, and family history on the sensitivity of first screening mammography. *JAMA* **276**(1): 33–38.

Kobrunner SH, Hacker A, Sedlacek S (2011) Advantages and disadvantages of mammography screening. *Breast Care* **6**: 199–207.

Kopans DB, Smith RA, Duffy SW (2011) Mammographic screening and 'overdiagnosis'. *Radiology* **260**: 616–620.

Kurian AW, Munoz DF, Rust P, Schackmann EA, Smith M, Clarke L, Mills MA, Plevritis SK (2012) Online tool to guide decisions for BRCA1/2 mutation carriers. *J Clin Oncol* **30**(5): 497–506.

Lagerlund M, Hedin A, Sparen P, Thurfjell E, Lambe M (2000) Attitudes, beliefs, and knowledge as predictors of nonattendance in a Swedish population-based mammography screening program. *Prev Med* **31**: 417–428.

Leonard GD, Swain SM (2004) Ductal carcinoma in situ, complexities and challenges. *J Natl Cancer Inst* **96**(12): 906–920.

Leong SPL, Shen ZZ, Liu TJ, Agarwal G, Tajima T, Paik NS, Sandelin K, Derossis A, Cody H, Foulkes WD (2010) Is breast cancer the same disease in Asian and Western countries. *World J Surg* **34**(10): 2308–2324.

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equation of state calculations by fast computing machines. *J Chem Phys* **21**: 1087–1092.

Narod SA (2012) Tumour size predicts long-term survival among women with lymph node-positive breast cancer. *Curr Oncol* **19**(5): 249–253.

Plevritis SK, Salzman P, Sigal BM, Glynn PW (2007) A natural history model of stage progression applied to breast cancer. *Stat Med* **26**: 581–595.

Plummer M, Best N, Cowles K, Vines K (2006) CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**(1): 7–11.

Poole D, Raftery AE (2000) Inference for deterministic simulation models: the Bayesian melding approach. *J Am Stat Assoc* **95**(452): 1244–1255.

Core Team R (2012) R: A Language and Environment for Statistical Computing.

Straatman H, Peer PG, Verbeek AL (1997) Estimating lead time and sensitivity in a screening program without estimating the incidence in the screened group. *Biometrics* **53**(1): 217–229.

Sun X, Faunce T (2008) Decision-analytical modelling in health-care economic evaluations. *Eur J Health Econ* **9**(4): 313–323.

Tabar L, Fagerberg G, Chen H, Duffy SW, Smart CR, Gad A, Smith RA (1995) Efficacy of breast cancer screening by age. new results from the Swedish two-county trial. *Cancer* **75**(10): 2507–2517.

Venables WN, Ripley BD (2002) *Modern Applied Statistics with S.* 4th edn. Springer: New York, USA.

Wu JC, Hakama M, Anttila A, Yen AM, Malila N, Sarkeala T, Auvinen A, Chiu SY, Chen H (2010) Estimation of natural history parameters of breast cancer based on non-randomized organized screening data: subsidiary analysis of effects of inter-screening interval, sensitivity, and attendance rate on reduction of advanced cancer. *Breast Cancer Res Treat* **122**: 553–566.

Zwillinger D (2011) *CRC Standard Mathematical Tables and Formulae.* 32 edn: (CRC Press).

Supplementary Information accompanies this paper on British Journal of Cancer website (http://www.nature.com/bjc)