European Journal of Epidemiology (2018) 33:465–472
https://doi.org/10.1007/s10654-017-0339-7

SCREENING

Disaggregating the mortality reductions due to cancer screening: model-based estimates from population-based data

James Anthony Hanley[1] · Sisse Helle Njor[2,3]

Received: 25 July 2017 / Accepted: 28 November 2017 / Published online: 5 December 2017

Abstract
The mortality impact in cancer screening trials and population programs is usually expressed as a single hazard ratio or percentage reduction. This measure ignores the number/spacing of rounds of screening, and the location in follow-up time of the averted deaths vis-a-vis the first and last screens. If screening works as intended, hazard ratios are a strong function of the two Lexis time-dimensions. We show how the number and timing of the rounds of screening can be included in a model that specifies what each round of screening accomplishes. We show how this model can be used to disaggregate the observed reductions (i.e., make them time-and screening-history specific), and to project the impact of other regimens. We use data on breast cancer screening to illustrate this model, which we had already described in technical terms in a statistical journal. Using the numbers of invitations different cohorts received, we fitted the model to the age- and follow-up-year-specific numbers of breast cancer deaths in Funen, Denmark. From November 1993 onwards, women aged 50–69 in Funen were invited to mammography screening every two years, while those in comparison regions were not. Under the proportional hazards model, the overall fitted hazard ratio was 0.82 (average reduction 18%). Using a (non-proportional-hazards) model that included the timing information, the fitted reductions ranged from 0 to 30%, being largest in those Lexis cells that had received the greatest number of invitations and where sufficient time had elapsed for the impacts to manifest. The reductions produced by cancer screening have been underestimated by inattention to their timing. By including the determinants of the hazard ratios in a regression-type model, the proposed approach provides a way to disaggregate the mortality reductions and project the reductions produced by other regimes/durations.

Keywords Screening, mortality, non-proportional hazards · Birth-cohorts · Lexis diagram · Disaggregation · Design matrix

Introduction

A single hazard ratio is appropriate if the reduction in hazard rates is immediate and sustained. Examples include the near-immediate and continued protection against HIV acquisition following adult circumcision, the decades of protection afforded by a vaccine, and the near immediate and sustained mortality reduction from one-time-screening for abdominal aortic aneurysms [1]. A single ratio is also appropriate if—as with blood thinners/beta-blockers—one limits the time-window to when the agent is active.

Cancer screening comparisons *generate non-proportional* hazards: mortality reductions appear after some delay following the first screen, and eventually disappear following the last one. In prostate cancer screening, the delay is considerable. After an average of 9 years [2] the reported hazard ratio (HR) was 0.8, i.e., the average reduction was 20%. However, hazard rates only began to diverge after 7 years; a re-analysis [3] using time-specific data made this delay even clearer. As one commentator [4] wrote, “Perhaps a better summary … is not the 20% overall reduction … but the combination of no reduction in the first

✉ James Anthony Hanley
james.hanley@McGill.CA

[1] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 1020 Pine Ave. West, Montréal, QC H3A 1A2, Canada

[2] Department of Public Health Programmes, Randers Regional Hospital, Randers, Denmark

[3] Department of Clinical Epidemiology, Aarhus University, Aarhus, Denmark

seven or so years and a reduction of about 50% after 10 years." Despite the reported substantial inter-country differences in the screening intervals (2 or 4 years), in the upper screening ages, and in the length of follow-up (3–19 years), the meta-analysis [5] ignored these critical HR determinants and averaged the 20% with ones from other studies with even more varied determinants [6]. The sparse data beyond 10 years made it impossible to judge if any HR had reached its nadir, let alone when after the last screen it reverted towards unity.

One recent report that did use non-proportional hazards—but not as the primary method of analysis—is the one on the ovarian cancer screening trial [7].

The 30-year follow-up of a colon cancer screening trial [8] was long enough to 'see' the HR curve revert, provided cumulative mortality is disaggregated [9]. The re-analysis [9], as in 2005 [10], showed the importance of the number and timing of screens, and the unplanned hiatus in screening—a peculiarity some commentators [11] ignore. Single HRs from overly-long follow-up (e.g. [12]) can contain substantial ''post screening noise'' [13], unless, as others have done [14], the analysis focuses only on those deaths that might have being averted by screening. Ones calculated when the single HR *first* becomes statistically significant from 1 are also problematic.

Policy regarding breast cancer screening continues to lean heavily on the data from older trials, many of which involved only a few rounds of screening. Moreover, as we documented [15], few of the primary analyses linked the numbers of fatal cancers back to the screening history in the relevant years before these deaths. The 'delay' principle [16] was first employed [17, 18] to re-analyse the one trial with extensive screening, but subsequent meta-analyses continue to ignore the variations in the numbers of screens, and the time windows in which mortality reductions would and would not be expected.

The purpose of this paper is to introduce epidemiologists to a statistical model which we have recently developed to describe the not-constant-over-time hazard ratios generated by cancer screening. The technical details, along with applications to colon and lung cancer screening trials, have already been published in a statistical journal [9], and in a doctoral thesis [19]. Our purpose here is to introduce it, and the basic principles behind it, to a broader audience; we illustrate it by applying it to the screening of a cancer for where there are no recent trials, but considerable population-level data, much of them from the 21st century.

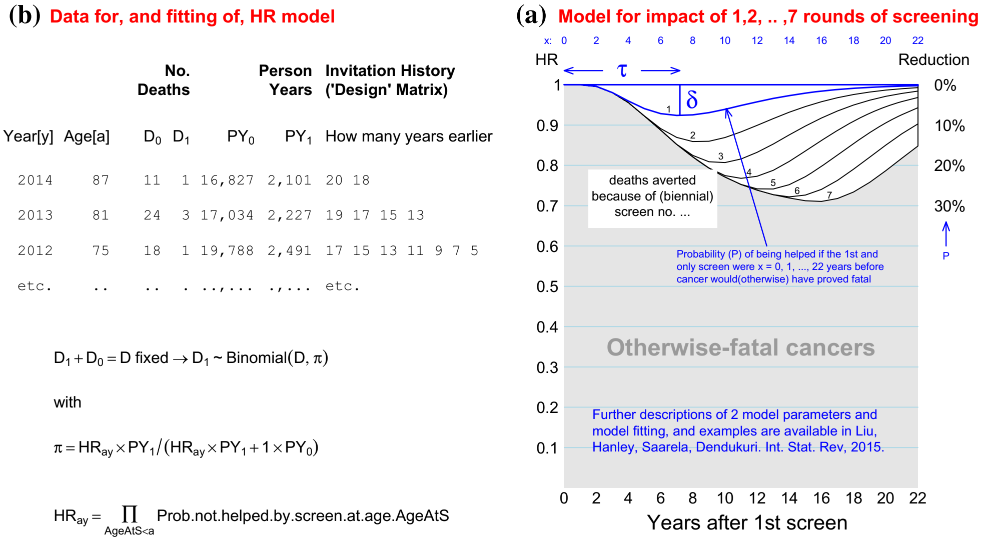
Indeed, the best contemporary evidence concerning the benefit of mammography comes from populations that introduced programs in phases. For example, two Danish areas, Funen and Copenhagen introduced screening well before the rest of Denmark; this allowed [20] comparisons of mortality rates over the next 10–14 years while using data from the preceding 10–14 years to adjust for inherent regional differences, and temporal improvements in treatment. Because of its larger numbers of deaths, and the similarity of the pre-screening mortality rates in the screened and non-screened areas, we illustrate the importance of the HR determinants using data from Funen [21]. From November 1993 onwards, women aged 50–69 in this region were invited to biennial mammography while those in most other regions of Denmark were not until the end of 2007. This article illustrates how the mortality reductions over the relevant {attained-age, calendar year} or 'Lexis' cells can be fitted as a function of 2 parameters and the invitation histories for these cells, and how the disaggregated (i.e. time-specific) effects of one round can be quantified and used for projections.

The model, the data, and the parameter-fitting

The model

The model has been described elsewhere [9, 19, 22, 23] and applied to data from colon and lung trials [9], so we give only a brief description. We do so by considering as an example women aged 50 in 1994, who, in the absence of a screening program, would subsequently be diagnosed with breast cancers that turned out to be fatal. For the as-yet-undiagnosed breast cancers that were to prove fatal at age $a = 52$ say, a *single* round of screening at age 50—some $x = 2$ years before it was fatal—would have a high probability of detecting and treating them earlier, but a low probability of effecting a cure. For those that would prove fatal at $a = 82$, the opposite would be the case. This detectability-curability trade-off means that the probability (P) that a woman whose cancer proved fatal at age 'a' would have been helped by a single round of screening at age $a - x$ is a strong function of x, the number of years since that screen. To begin with, we adopt a simple functional form, shown in blue in Fig. 1a. The probability attains its maximum, say δ, at some time τ, and falls monotonically on both sides of this. These two parameters, to be estimated from the observed data, are the essence of the model.

What if women—who in the absence of screening were to die of breast cancer—had (been invited to) *several* screens, say every 2 years from age 50 until 69? The reduction in the number of cancer deaths at age a is now an aggregate of the (staggered) contributions of all of the rounds up to then; the resulting bathtub shaped hazard ratio function is shown in black in Fig. 1a, and its value at a particular age represents the proportion of otherwise-fatal



		No. Deaths		Person Years		Invitation History ('Design' Matrix)
Year[y]	Age[a]	D_0	D_1	PY_0	PY_1	How many years earlier
2014	87	11	1	16,827	2,101	20 18
2013	81	24	3	17,034	2,227	19 17 15 13
2012	75	18	1	19,788	2,491	17 15 13 11 9 7 5
etc.	..	..	.	..,...	.,...	etc.

$D_1 + D_0 = D$ fixed → $D_1 \sim \text{Binomial}(D, \pi)$

with

$\pi = HR_{ay} \times PY_1 / (HR_{ay} \times PY_1 + 1 \times PY_0)$

$HR_{ay} = \prod_{AgeAtS<a} \text{Prob.not.helped.by.screen.at.age.AgeAtS}$

Fig. 1 Schematic showing the model for the reductions produced by one or more rounds of screening, the required data to fit the 2 parameters δ and τ, and the fitting of these two parameters. Shown in blue in panel **a** is the probability (P) that cancers that (in the absence of screening) proved fatal at age *a* would have been averted by the possibly earlier treatment prompted by a single round of screening *x* years earlier. *x* is shown in blue along the horizontal axis at the top. As shown by the blue arrow, it is approximately 6% when $x = 10$ years. The probability is greatest, at δ percent, when the screen was τ years previously. Shown as black, again as a function of *x*, are the probabilities (P) that these otherwise fatal cancers would have been averted as a result of 2, 3, … 7 rounds of screening offered every two years from age $a - x$ onwards, where *x* denotes the length of time between the first screen and attaining age a. The complement of P[*x*] can be interpreted as the probability that, despite screening, the cancer will still prove fatal. It can also be interpreted as a Hazard Ratio (HR) at age *a* that is ≤ 1. The proportion (probability) itself can be interpreted as the reduction in the mortality rate at age *a* in persons for whom it has been *x* years since their first screen (horizontal axis at bottom). Compared with the single-round HR in blue, the HR generated by multiple screens extends deeper, over a longer time-window, and exhibits a bathtub shape with a delay, a nadir or sustained asymptote, and an eventual return to 1 after all the effects of the last screen have been expressed. Shown in panel **b** are the data for, and fitting of the 2 parameters (δ and τ) of the model. (Color figure online)

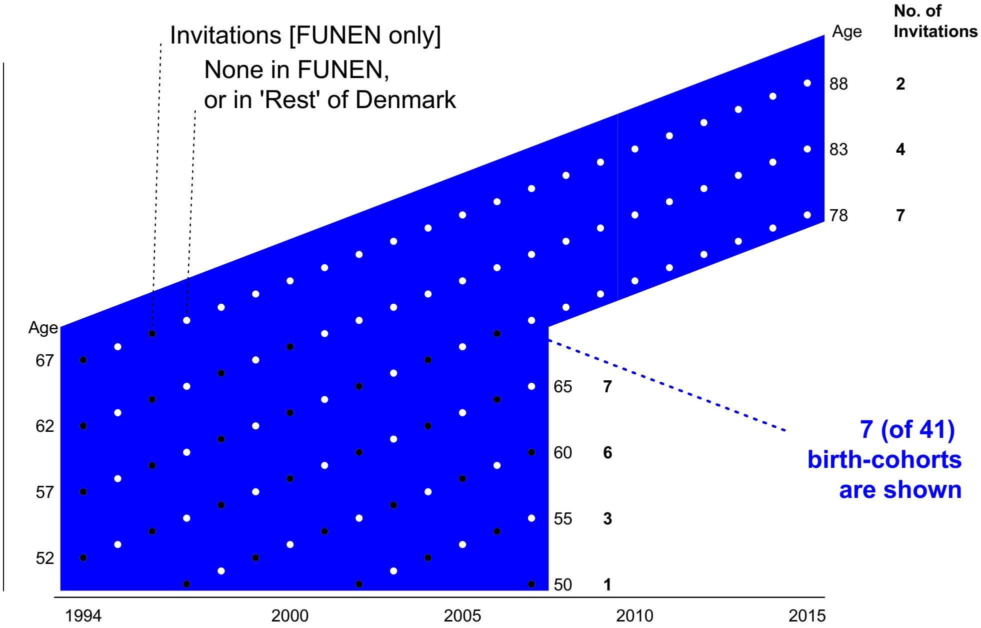
cancers at that age that would still be fatal *despite* the screening.

Like other trials/programs, Funen did not limit the invitations to one age (50) in one year (1994). It invited all birth cohorts every 2 years while they are between age 50 and 69. The invitations can be visualized in what is known as a 'Lexis Diagram' [24], which shows how different cohorts progress simultaneously along the two time scales of age—on the vertical axis—and calendar time—on the horizontal axis. In the data-analysis, we will divide the ages and years into 1-year bins that taken together form small 1 × 1 Lexis 'squares' or 'cells,' and use the number of breast cancer deaths in each small square in each region as a separate Poisson random variable. Thus, as is seen in the Lexis Diagram in Fig. 2, those oldest when the program was begun, and youngest at the last invitation before the follow-up ended, did not receive as many invitations as those who are 50 when the program started. As a result of these variations, and of the 'delay' principle', the HR 'surface' over this Lexis space must be a strong function of the age and calendar-year (or age and follow-up year) time scales.

The data

We retrieved data from the Danish cause of deaths register on all breast cancer deaths until 31 December 2015. Data on invitation to mammography screening in Funen were retrieved from the Funen mammography screening register. For each of the relevant ages (*a*) in each of the 22 years (*y*) after the Funen program began, the data consisted of the numbers of breast cancer deaths (D_1 and D_0), and corresponding women years (WY_1 and WY_0), in Funen ($_1$) and the parts of Denmark where mammography screening did not start until late 2007 (RestDK) ($_0$). The values for 3 selected cells are shown in the rows in panel (b) of Fig. 1, along with when—counting back from (a,y)—the Funen birth cohort received screening invitations. These screening histories can be thought of as the 'Design Matrix' in this regression-type model. Since the breast cancer mortality rates in the years before 1994 were very similar in Funen and the comparison region, we ignore these pre-screening data. The original Njor article also documented the degree of opportunistic screening, breast cancer treatment protocols, and multidisciplinary breast cancer management teams in Funen before and during screening, and in the rest of Denmark in the same calendar periods. As was done in

Fig. 2 Schematic of the screening invitations extended to, and follow-up of, women in Funen birth cohorts (7 shown). None were extended to the corresponding cohorts of women in the "rest' of Denmark until late 2007



both the Copenhagen and Funen studies, a background difference could also be accommodated by including the pre-1994 data, and by including in our model a parameter representing this difference.

The fitting

Figure 1b shows data for three selected (*a*,*y*) Lexis cells, with PY_1 and PY_0 person years in the invited and uninvited, and numbers of deaths D_1 and D_0. If the latter are assumed to follow two Poisson distributions, and if one conditions on $D = D_1 + D_0$, then $D_1 \mid D$ follows a binomial distribution with 'denominator' D and a 'proportion' parameter π that is a function not just of PY_1 and PY_0, but also of how 'non-null' the hazard ratio is at that point in time [24]. For example, in the third row of Fig. 1b, if the HR were 0.8, then the expected split of the 19 deaths should be proportional to (2491 × 0.8): (19,788 × 1), or 1.7:17.3, yielding a Binomial distribution with '*n*' = 19 and π = 0.09. The hazard ratio HR[*a,y*] [9, 19, 24] in cell (a,y) is a function of the two model parameters (δ,τ) and the number and timing of the preceding screening invitations. Since the HR in a cell also represents the proportion of otherwise-fatal cancers that would still be fatal *despite* the screening, it was calculated as the probability that each of the preceding rounds of screening failed to avert the death, i.e. as the product of the complements of the P function described above, evaluated at the time-lags corresponding to these preceding rounds. See the last equation in Fig. 1b and the convolutions pictured in Fig. 1a. As explained elsewhere [9, 19], the probability function was taken to have a gamma function shape, but with the scale parameter constrained (larger amounts of data would have allowed this constraint to be removed). The two model parameters δ and τ were fitted by summing the cell-specific log-likelihood contributions, and numerically maximizing the sum.

Results

Over all ages and follow-up years in the Lexis diagram, the 'average' Funen-RestDK difference, i.e., the 'reduction' or 'deficit' in breast cancer mortality in Funen that is 'attributable' to the screening, was 18%. This is a smaller reduction that the 22% seen in the follow-up that ended on December 31, 2009 [21]. Part of this difference may be the play of chance, and part may be because we now include deaths from cancers that are only diagnosed after the women stopped being screened (at age 70).

To motivate the model-based measures, we first present year-specific comparisons in Fig. 3. Once segregated into 3 birth cohorts, each 5 years wide, the yearly numbers of deaths in Funen are in the single digits, and so the year-specific mortality rate differences are noisy. With the help of some smoothing, however, it seems that the reductions in those who—because they were already in their late 60s in 1994—received the fewest invitations (red) do not persist for as long as those in the cohorts—in their late 50s in 1994—who received the most (blue). Moreover, the reductions in the intermediate (green) cohorts—in their early 60s in 1994—also began to disappear earlier.

The model-based estimates were that the maximum probability of being helped by a single round of screening

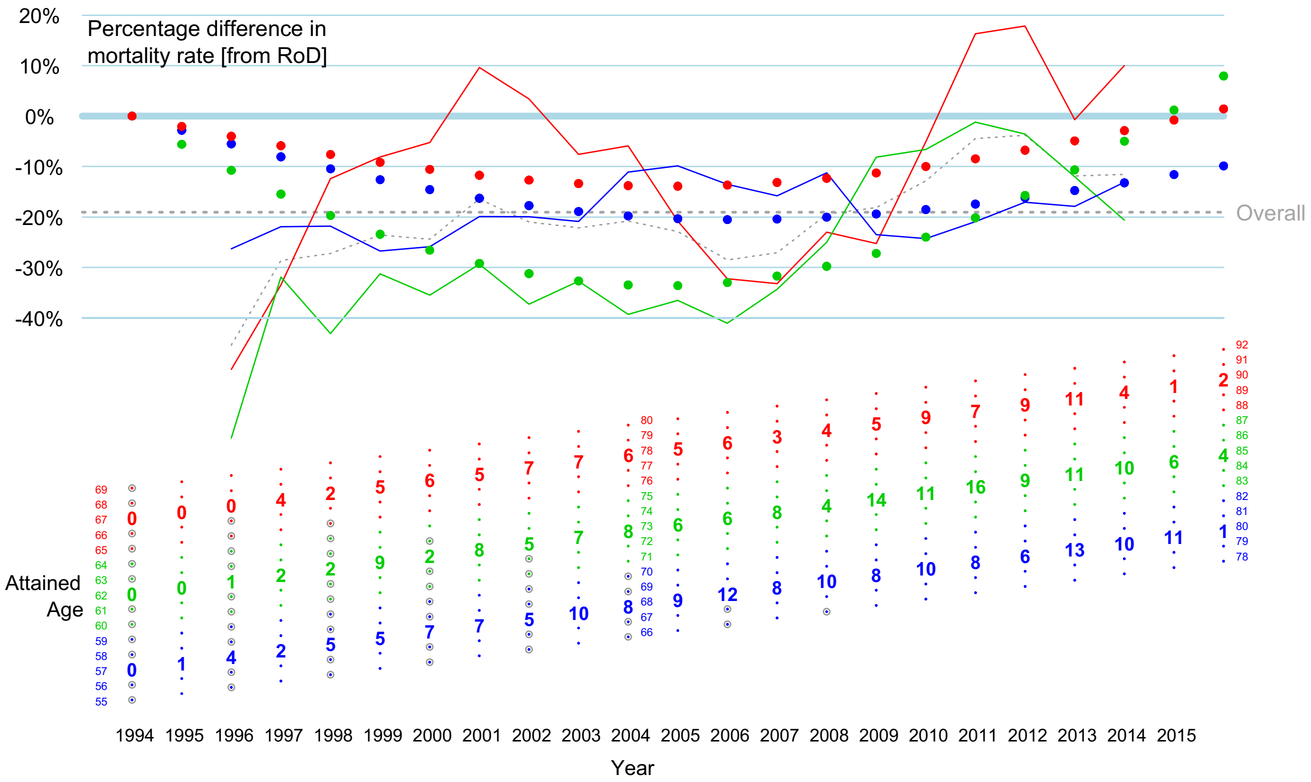


Fig. 3 Average, and followup-year-specific, differences in breast cancer mortality, in 3 birth cohorts, each 5 years wide (color-coded), together with yearly numbers of breast cancer deaths in the Funen cohorts [The rest of Denmark has approximately 8 times more women-years than Funen]. In the modified Lexis diagram in the bottom panel, grey circles indicate invitations to those Funen women who attained the indicated ages in the years indicated. Numbers are numbers of deaths from breast cancer in the 3 age-bands. Percentage differences in upper panel: Dotted line: age-year-matched Mantel–Haenszel ‘average’, 3 lines: age-matched Mantel–Haenszel year-specific, 3 smooth patterns: cohort-specific natural cubic splines, each with 2 degrees of freedom. (Color figure online)

was 8% (δ = 0.08), at τ = 8 years [values close to these were used to draw Fig. 1a]. Thus, for women who, in the absence of screening, died at age *a*, the optimum timing of a *single* screen would have been at age *a* − 8. The fitted reductions obtained by coupling these two parameter values with the ‘invitation histories’ for each of the Lexis cells are shown in Fig. 4. The smallest yearly ‘dividends’, ranging from 0 to 8% to 0%, were in those aged 68–69 in 1993–1994 received just 1 invitation. Among Funen women invited for their first mammography screening at age 62–63 and invited four times before turning 70, mammography screening averted 2% of the deaths that would otherwise have occurred in year 1996; 23% of deaths in year 2005 and 3% of deaths in year 2015. The largest reduction of 30% was seen among the (youngest) cohort who received the most invitations, and could be followed until 2015. This cohort was aged 54–55 when mammography screening began.

The 0% fitted reductions in year 1 should not be taken literally, since there is rounding involved, and there may also have been some delay in implementing the first round of screening. In incidence based studies, there are fewer cancer deaths in the first follow-up years than in the latter ones, a feature that several cancer screening trials did not include in their precision/power planning. [Because of the very long accrual period and relatively short follow-up of the ERSPC, the distribution of cancer deaths did not reflect this: those in the early follow-up years outweighed those in the later years, and unduly weighted the average mortality reduction towards the minimal reductions seen in the earlier ones [3].]

Discussion

Recently, the long-recognized [25] but typically-ignored ‘delay’ principle was forcefully stated: “the proportional reduction in mortality from the cancer is nothing like a constant over time from the beginning of the screening to the end of the follow-up (for an arbitrary duration of it)” [16]. Until now, reports on breast cancer mortality reductions lacked models to address this non-constant reduction, and had to rely on a single HR [6]. Thus, they implicitly

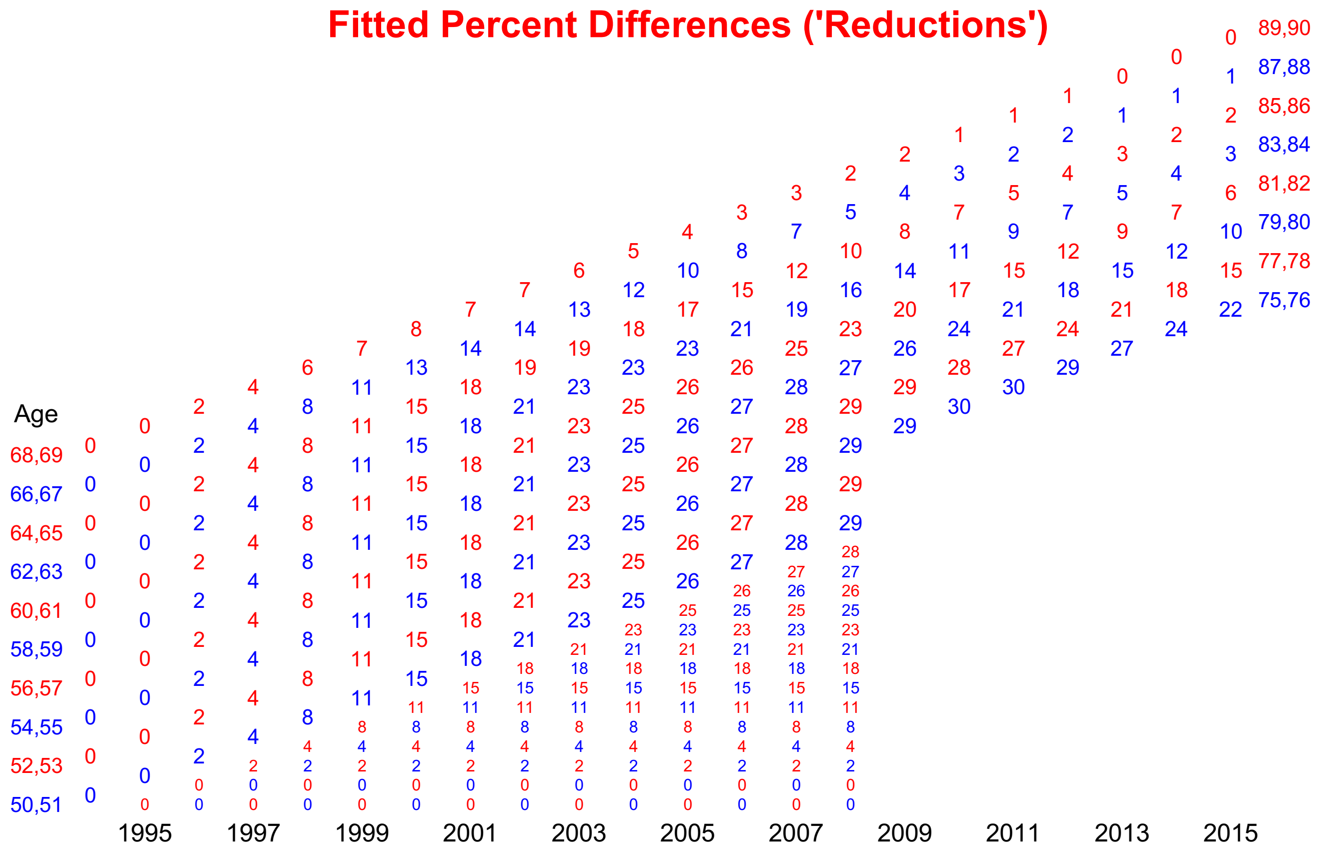


Fig. 4 For each birth cohort, the age-and year-specific fitted percentage reductions in breast cancer mortality. They were derived from the Maximum Likelihood estimates of the two model parameters (maximum probability of being helped by a single round of screening 8 years previously: 9%) and the number and timing of the preceding screening invitations

assumed a proportional hazards model where reductions are constant over follow-up time.

The proposed model is a first step towards describing the time-specific reductions a sustained screening program might produce. Whereas earlier efforts used moving averages [18], or directly fitted a smooth HR curve [3] without regard to the screening schedule, the present approach uses fundamental (rather than *design*-dependent) parameters that, coupled with the schedule (the design matrix), produce a HR *function*.

The average 18% reduction one obtains either by fitting a proportional hazards model over the Lexis cells, or using them as strata in a Mantel–Haenszel summary ratio, does not mean that 10 biennial screenings from 50 to 69 would avert 18% of the breast cancer deaths that would otherwise have occurred. This single estimate is arbitrary, and particular to the age-mix at intake, the numbers of invitations received, and duration of follow-up. The model-based cell-specific reductions are much more realistic, and show what was accomplished by the various amounts of screening up to the ages and years in question. As expected, the reductions vary considerably in age and time: cohorts first screened in their 50s—and thus more often—had much larger mortality deficits that those first screened at later ages—and less often.

Our proposed model separates the fundamental 'screening ability' parameters (δ,τ) from the design matrix (each row of which is the invitation history for a Lexis cell); thus, as in a regression context, it allows one to estimate the HR curve for a new 'row,' i.e. a specific screening frequency and duration. The overall 18% reduction, and the single-percentage reductions reported from all screening trials do not correspond to any specific estimand, but rather to an average over some mix of frequencies and durations, and follow-up years.

Traditionally, cost–benefit models of a sustained screening program have been quite complex. The disaggregated reductions derived from our approach, coupled with the desired screening schedule, provide a transparent yet flexible way to project the benefits with screening regimes that have not been tested. As an unusual but telling example, the average reduction of 22% in the biennial screening arm of the colon cancer screening trial [8] was computed over 30 years without considering the number of screens or the substantial but unplanned screening hiatus. The model allowed us to 'fill in' the missing screens and compute the (larger) reduction that would have been realized with the intended regime [9]. We used the same compliance rate (78%) as was observed in the trial, but, as explained in the thesis, the HR function is readily modified to estimate what the percentage reduction curves would have been with a different rate, or with rates that vary from round to round. If compliance factors are not included in the HR function, the fitted parameter values produce an 'effectiveness' estimate; if they are, they can produce an 'efficacy' estimate.

In Funen, invitations to mammography were every 2 years until age 69, but the parameter estimates could be coupled with other scenarios, such as if invitations were discontinued at 65, or continued to age 75.

The use of a design matrix also means that we can accommodate data from different screening programs, where the invitation histories for the same {age, year} cell might differ.

The most common objection thus far to the model has been that the first and the subsequent rounds of screening are assumed to have the same impact. But, if there are sufficient data, the model is easily extended to allow separate parameter values for the first round, or to have them change smoothly with each additional round; with sufficient data it is also easy to substitute more complex forms of the probability function [9].

Another objection is that we used the same 2 parameters no matter the age at first invitation. Again, the issue is more with sparse data than with the model per se: with enough data, we could use separate parameter values for screens that begin at different ages. It was only our lack of data (see bottom of Fig. 3) that forced us to use common parameter values. Screening programs usually begin 'mid-stream' with a mix of 50-year olds who will eventually receive the full program, and older ones who have not yet had (*but in steady state would have had*) any screening. The parameters of ultimate concern are those that determine the steady state, but our estimates are weighted by a fatality mix that excludes early deaths in women for whom screening was not available when they were 50.

Another indication of the limits of our data is the limited resolution of goodness-of-fit statistics. We computed fitted numbers of breast cancer deaths in Funen, using six 'bins' that grouped adjacent birth cohorts. Against observed numbers of O = 165, 151, 123, 82, 37, and 10, the sum of the six {(O-Fitted)^2/Fitted} deviations was 3.3 for our 2-parameter model derived from screening principles, and 5.0 for the biologically unjustifiable proportional hazards model. This limited ability to distinguish between a biologically-based and a purely mathematical model was also evident in our numerical investigations [26].

Because of these data limitations, we are appealing to those in other countries who hold similar data to fit the proposed model to their data. Since the log-likelihood contributions are at the Lexis cell level, it is easy to aggregate data generated with different schedules or lengths of follow up etc. Thus we also invite data-holders to collaborate and contribute to a combined dataset that would have more 'reach' and that would allow us to test the model more extensively: for example, could the model fitted to data from an every-2-years regimen correctly predict the HR curve for say an every-3-years regimen? The most important feature of the model is the accompanying design matrix, which used uses the numbers of deaths *and the screening history* in each Lexis cell as the unit of analysis, rather than—by ignoring the history—treating the observations from these cells as exchangeable. Of particular value are Lexis cells where the effects have begun to disappear: it is not possible to reliably fit a bathtub shaped model without having data at the distal end. Because of the nature of cancer screening, there will be very few deaths at the beginning of the curve, so the HR values at that end are less critical.

Just as with any 'intention to' analyses, the HRs from our model do not distinguish between inherent limitations of the screening technique/biological model and lack of participation. However, different participation rates for different rounds (or countries) can be included in the binomial probabilities, thereby providing estimates of what the results would be with greater or lesser participation [9].

The model provides a natural way to handle the variations in studies that up to now have been considered exchangeable in meta-analysis. The critical requirements, we re-iterate, are the availability of both mortality data and invitation data *at the Lexis cell level*. By using these histories (instead of ignoring them or merely calculating a heterogeneity index) we can disaggregate the overall mortality reduction and help provide answers to questions such as: what is the optimal starting and stopping age? what is the optimal number of screenings?

Observational data, such as we have analyzed, can only be used if two comparable groups, where one group was invited to screening and the other was not, are available. In some countries it is not possible to find such groups, as the entire female population in a specific age group is invited. It is possible in countries where mammography screening was implemented in a staggered way. The time span between invitation of the two groups determines for how long the bathtub shape can be observed. From a statistical viewpoint, the optimal is to be able to observe the reduction until is returns to zero.

Acknowledgements The work was supported by a grant "Measuring the mortality reductions produced by cancer screening" to James

Hanley from the Canadian Institutes of Health Research (CIHR). CIHR had no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication; i.e., these activities were carried out by the authors, independently of the funder. Both authors had access to the Lexis-cell-level data used in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Authors' contribution James Hanley suggested the re-analysis; the authors jointly presented preliminary results, based on data to 2009, at the International Biometric Conference in Victoria, British Columbia, Canada in July 2016. Both drafted the manuscript, and both are guarantors for the study.

Compliance with ethical standards

Conflicts of interests Both authors declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval The study was approved by Region Midt under their umbrella permission form the Danish Data Inspection Board. According to Danish legislation this serves as ethical approval of register-based research projects (journal number 1-16-02-90-17).

References

1. Thompson SG, Ashton HA, Gao L, Scott RA, On behalf of the Multicentre Aneurysm Screening Study Group. Screening men for abdominal aortic aneurysm: 10 year mortality and cost effectiveness results from the randomised Multicentre Aneurysm Screening Study. BMJ. 2009;338:b2307. https://doi.org/10.1136/bmj.b2307.
2. Schröder FH, Hugosson J, Roobol MJ, et al. Screening and prostate-cancer mortality in a randomized European study. N Engl J Med. 2009;360:1320–8.
3. Hanley JA. Mortality reductions produced by sustained prostate cancer screening have been underestimated. J Med Screen. 2010;17(3):147–51. https://doi.org/10.1258/jms.2010.010005.
4. Law M. What now on screening for prostate cancer?. J Med Screen. 2009;16:109–11.
5. Djulbegovic M, Beyth RJ, Neuberger MM, Stoffs TL, Vieweg J, Djulbegovic B, Dahm P. Screening for prostate cancer: systematic review and meta- analysis of randomised controlled trials. BMJ. 2010;341:c4543. https://doi.org/10.1136/bmj.c4543.
6. Hanley JA. Measuring mortality reductions in cancer screening trials. Theme issue on screening. Epidemiol Rev. 2011;33:36–45. https://doi.org/10.1093/epirev/mxq021 **Epub 2011 May 30**.
7. Jacobs IJ, Menon U, et al. Ovarian cancer screening and mortality in the UK collaborative trial of ovarian cancer screening (UKCTOCS): a randomised controlled trial. www.thelancet.com. Published Online 17 Dec 2015. http://dx.doi.org/10.1016/S0140-6736(15)01224-6.
8. Shaukat A, Mongin SJ, Geisser MS, Lederle FA, Bond JH, Mandel JS, Church TR. Long-term mortality after screening for colorectal cancer. New Engl J Med. 2013;369:1106–14.
9. Liu Z, Hanley JA, Saarela O, Dendukuri N. A conditional approach to measure mortality reductions due to cancer screening. Int Stat Rev. 2015;. https://doi.org/10.1111/insr.12088.
10. Hanley JA. Analysis of mortality data from cancer screening studies: looking in the right window. Epidemiology. 2005;16(6):786–90.
11. Welch HG, Robertson DJ. Colorectal cancer on the decline—Why screening can't explain it all. N Engl J Med 2016;374:1605–7. https://doi.org/10.1056/NEJMp1600448A. 28 April 2016, see also: Kolata G, Medical mystery of the best kind: major diseases are in decline. New York Times, 8 July 2016.
12. Marcus PM, Bergstralh EJ, Fagerstrom RM, et al. Lung cancer mortality in the Mayo Lung project: impact of extended follow-up. J Natl Cancer Inst. 2000;92:1308–16.
13. Baker S, Kramer BS, Prorok PC. Early reporting for cancer screening trials. J Med Screen. 2008;15:122–9.
14. Tabár L, Vitak B, Chen TH, Yen AM, Cohen A, Tot T, Chiu SY, Chen SL, Fann JC, Rosell J, Fohlin H, Smith RA, Duffy SW. Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. Radiology. 2011;260:658–63.
15. Hanley JA. Analysis of mortality data from cancer screening studies: looking in the right window. Epidemiology. 2005;16(6):786–90.
16. Miettinen OS, Karp I. Epidemiological research: an introduction. Dordrecht: Springer; 2012.
17. Caro J, McGregor M. Screening for breast cancer in women aged 40–49 years. Montreal, Quebec, Canada: Agence d'évaluation des technologies et des modes d'intervention en santé [AETMIS, the Québec government agency responsible for health services and technology assessment]; 1993:91. (CETS report no. 22). http://www.aetmis.gouv.qc.ca/site/download.php?f1/4503b634ef04a597215ff3dc734d8d84e. Accessed 6 July 2005.
18. Miettinen OS, Henschke CI, Pasmantier MW, et al. Mammographic screening: no reliable supporting evidence? Lancet. 2002;359(9304):404–5. A fuller account can be found at http://image.thelancet.com/extras/1093web.pdf. Accessed 6 July 2005.
19. Liu Z. Measuring the mortality reductions due to cancer screening. Ph.D. Dissertation. McGill University Libraries. 2015. http://digitool.library.mcgill.ca/R/?func=dbin-jump-full&object_id=130321.
20. Olsen AH, Njor SH, Vejborg I, Schwartz W, Dalgaard P, Jensen M-J, Tange UB, Blichert-Toft M, Rank F, Mouridsen H, Lynge E. Breast cancer mortality in Copenhagen after introduction of mammography screening: cohort study. BMJ. https://doi.org/10.1136/bmj.38313.639236.82. Published 13 Jan 2005.
21. Njor SH, Schwartz W, Blichert-Toft M, Lynge E. Decline in breast cancer mortality: How much is attributable to screening? J Med Screen. 2015;22(1):20–7. https://doi.org/10.1177/0969141314563632msc.sagepub.com.
22. Hanley JA. Analysis of mortality data from cancer screening studies: looking in the right window. Epidemiology. 2005;16(6):786–90.
23. Liu Z, Hanley JA, Strumpf EC. Projecting the yearly mortality reductions due to a cancer screening program. J Med Screen. 2013;20:156–64. https://doi.org/10.1177/0969141313504088.
24. Clayton D, Hills M. Statistical models in epidemiology. New York: Oxford University Press; 1993.
25. Morrison AS. Screening in chronic disease. New York: Oxford University Press; 1985.
26. Hanley JA, Liu Z, Saarela O. Fitting a model of the mortality reductions produced by one/several rounds of cancer screening: time and sample size considerations. Presentation (and abstract). In: 2016 Meeting of the Statistical Society of Canada, Brock University, Ontario, Canada.

European Journal of Epidemiology (2018) 33:433–435
https://doi.org/10.1007/s10654-018-0406-8

COMMENTARY

Statistical analysis and decision making in cancer screening

Dik Habbema[1]

Received: 30 April 2018 / Accepted: 3 May 2018 / Published online: 12 May 2018

Randomized controlled trials are the primary source of evidence for assessing the effectiveness of cancer screening. Thus far, trial data have mainly been analyzed using relative risk estimates or proportional hazard models [1]. Proportional hazard models assume that screening results in a time-independent reduction in cancer mortality. Hanley, Liu and coworkers have developed a model with a time-dependent mortality reduction, see elsewhere in this issue of the journal [2, 3]. The model assumes that the reduction in mortality from the target cancer appears after a delay following a screen, and eventually disappears. Mortality reductions from subsequent screening rounds are superimposed. The resulting function has a bathtub form, and is determined by two parameters: the time between a screen and maximum relative mortality reduction, and the value of the maximum relative mortality reduction [2]. The authors have applied their method to data from prostate cancer [1], lung cancer [3], colorectal cancer [3] and breast cancer [2], using excellent graphical illustrations (Figures 3 and 4 in [3]).

The Hanley–Liu model is more realistic than the proportional hazard model. In practice, discriminating between the two models can be difficult. Designers of screening trials are aware of the bathtub dynamics of mortality reduction. They mitigate the influence of the initial phase of (near) absence of reduction by excluding persons with an already established diagnosis of the target cancer. A good compromise follow-up duration is the crux for dealing with the tapering off phase at the end. Follow-up should neither be too short when mortality reduction is still increasing nor too long with much noise from deaths which could not have been prevented by screening anyhow. With these choices, most cancer deaths in screening trials will occur in the bottom part of the bathtub, where the constant mortality reduction of the proportional hazard model is a good approximation to the Hanley–Liu model. And indeed, it proved not to be possible to discriminate between the two models in the analysis of the Danish breast cancer data [2]. The scatter of the time-dependent relative mortality dots in Figures 3 and 4 in [3] suggests that this might also be the case for the lung cancer and colorectal cancer analyses. This lack of discrimination with more complex models might be a reason why the simple proportional hazard model has persisted as the model of choice for statistical analysis of trial data.

The time-dependent mortality reduction curve of the Hanley–Liu model allows us to reflect on trial design issues like screening interval, follow-up time and power analysis.

In order to provide maximal information, the interval between subsequent screenings should be sufficiently long to provide information about the whole trajectory of the bathtub mortality reduction curve. A trial with 3-year intervals will be more informative than a trial with 1-year intervals.

Contrary to the proportional hazard model, duration of follow-up is not crucial for the Hanley–Liu model. While mortality after long follow-up is a source of random noise in the proportional hazard model, it is informative in the Hanley–Liu model for estimating the dynamics of the mortality reduction.

The high costs of screening trials strongly depend on their size. Because of the use of the time-dimension of the mortality data, power calculations will undoubtedly lead to a smaller sample size for the Hanley–Liu model than for the proportional hazard model.

Hanley and Liu note that use of their model is hindered by sparse data. This problem would even become worse when important determinants like age at first invitation and rank of the screening round would be included in the model [2]. The appeal of Hanley and Liu to screening data owners to collaborate is therefore timely and should be endorsed. In addition, it could be recommended that Lexis diagrams as used by Hanley and Liu, with number of deaths and person years at risk in each cell [2] should routinely be included in reports of screening trial results. The Lexis diagram has an age- and a calendar-time axis, describes

✉ Dik Habbema
j.d.f.habbema@erasmusmc.nl

[1] Department of Public Health, Erasmus MC University Medical Center, Rotterdam, The Netherlands

how cohorts progress along these axes and constitutes a database for further epidemiologic analysis [4].

Mortality analysis of screening trials usually takes place between 15 and 30 years after start of the trial. During this period, some of the biological and behavioral processes which underlie the mortality effects of cancer screening will have changed. Underlying processes which can change over time include incidence of cancer, the stage distribution of diagnosed cancers in the absence of screening, the stage-specific survival of cancer with current treatment, the sensitivity and specificity of screening tests in different disease stages, compliance to the screening, the characteristics of further diagnostics in case of an abnormal screening test result, and the stage specific survival in screen detected cancers, including precursor lesions. For example participants in the Minnesota trial for colorectal cancer screening were (healthy) volunteers, and since the trial the FOBT has largely been replaced by quantitative immunochemical blood tests and new cancer treatments have become available. The proportional hazards and Hanley–Liu models can both be characterized as modeling the mortality response to a screening stimulus which is delivered in the context of underlying processes. The models have no mechanism for correcting the response for secular changes in the underlying processes. This is a major problem for using the results of a statistical analysis beyond the trial context, for example for guideline development.

Many beneficial and harmful outcomes have to be taken into account when comparing screening policies, including overtreatment, anxiety after positive screening tests and complications from screening, follow-up tests and treatment. See [5] for a table of outcomes for colorectal cancer screening. Only one of the outcomes, mortality, is addressed by the proportional hazards and Hanley–Liu models. Mortality is arguable the most important outcome, as cancer screening without mortality reduction is useless.

The mortality output of the Hanley–Liu model which consists of the curve of relative mortality between screening and control group has to be processed before it can be used in decision making. A switch has to be made from relative to absolute mortality, in order to avoid that high and low cancer incidence situations would be treated the same. Age of death should be taken into account by calculating the expected number of life-years gained when preventing a death. Otherwise, prevented deaths at age 50 and age 90 would be valued the same. Two further possible actions are adjustment for time-preference by putting more weight on nearby compared to far away life-years, and adjustment for quality of life by calculating quality-adjusted life years [6].

The suggestion that the Hanley–Liu model can be used for deriving optimal ages and frequency of screening [2] is rather optimistic in view of the need to correct for secular changes and to weigh many harms and benefits. It might be better to turn to mathematical models which are developed with their use for decision making in mind. These so-called decision analytic models consider demography, epidemiology, natural history, screening tests, treatment and other processes, and aim to integrate available data to estimate the health consequences of alternative screening strategies [7]. By now, decision analytic models have been developed in many fields of medicine. For cancer screening, a large number of model groups collaborate in the Cancer Intervention and Surveillance Modeling Network (CISNET). The models have been described in a standardized way, see https://cisnet.cancer.gov/resources/profiles.html. Decision analytic models are increasingly used for informing screening guidelines development, for example by the United States Preventive Services Task Force [8, 9].

The scientific status of decision analytic models is unclear. While statistical models are developed within the firm context of probability theory and theoretical statistics [3], relevance is the primary concern in the development of decision analytic models. In order to increase their trustworthiness, general recommendations for good research practice in decision analytic modeling have been formulated [7]. For cancer screening, model quality and relevance have been discussed in [10]. The quality and credibility of decision models strongly depends on their performance in reproducing results of screening studies. They are considered most useful in situations where strong primary data are available [10]. For example, parameters of a decision analytic model for colorectal cancer screening could be fitted to the results of three randomized trials [11]. In view of the complexity of decision analytic models, much can be gained from collaboration between modeling groups [12] and from multi-model studies [13].

In conclusion, statistical models and decision analytic models are both important in cancer screening. Statistical models are essential for analysis of trial data. Decision analytic models are used in screening guidelines development. Decision modelers can learn from statistical models for improving the fitting and validation of primary data. Statistical modelers can learn from decision analytic models for improving the usefulness of their models for decision making. Hanley and Liu have improved on existing statistical models. By modeling the time dimension of the mortality reduction they improved the relevance for decision making, especially with regard to the question of optimal screening intervals. Decision analytic modelers should in turn try to learn from the Hanley–Liu model for improving the ways in which they fit their model to primary data.

References

1. Hanley JA. Measuring mortality reductions in cancer screening trials. Epidemiol Rev. 2011;33:36–45.
2. Hanley JA, Njor SH. Disaggregating the mortality reductions due to cancer screening: model-based estimates from population-based data. Eur J Epidemiol. 2017. https://doi.org/10.1007/s10654-017-0339-7.
3. Liu Z, Hanley JA, Saarela O, Dendukuri N. A conditional approach to measure mortality reductions due to cancer screening. Int Stat Rev. 2015. https://doi.org/10.1111/insr.12088.
4. Clayton D, Hills M. Statistical models in epidemiology. New York: Oxford University Press; 1993.
5. Harris R. Speaking for the evidence: colonoscopy vs computed tomographic colonography (editorial). J Natl Cancer Inst. 2010;102:1212–4.
6. Gold MR, Siegel JE, Russel LB, Weinstein MC. Cost-effectiveness in health and medicine. Oxford: Oxford University Press; 1996.
7. Medical Decision Making. Special Issue: Recommendations of the ISPOR-SMDM Joint Modeling Good Research Practices Task Force. 2012;32:667–743.
8. Zauber AG, Lansdorp-Vogelaar I, Knudsen AB, Wilschut J, van Ballegooijen M, Kuntz KM. Evaluating test strategies for colorectal cancer screening: a decision analysis for the U.S. Preventive Services Task Force. Ann Intern Med. 2008;149:659–69.
9. Mandelblatt JS, Cronin KA, Bailey S, Berry DA, de Koning HJ, Draisma G, et al. Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. Ann Intern Med. 2009;151:738–47.
10. Habbema JDF, Wilt JW, Etzioni R, Nelson HD, Schechter CB, Lawrence WF, Melnikow J, Kuntz KM, Owens DK, Feuer EJ. Models in the development of clinical practice guidelines. Ann Intern Med. 2014;161:812–8.
11. Lansdorf-Vogelaar I, van Ballegooijen M, Boer R, Zauber A, Habbema JDF. A novel hypothesis on the sensitivity of the fecal occult blood test. Results of a joint analysis of 3 randomized controlled trials. Cancer. 2009;115:2410–9.
12. Habbema JDF, Schechter CB, Cronin KA, Clarke LD, Feuer EJ. Modeling cancer natural history, epidemiology and control: reflections on the CISNET breast group experience. J Natl Cancer Inst Monogr. 2006;36:122–6.
13. Berry DA, Cronin KA, et al. Effect of screening and adjuvant therapy on mortality from breast cancer. N Engl J Med. 2005;353:1784–92.