The Cost-Effectiveness of Screening Programs Using Single and Multiple Birth Cohort Simulations: A Comparison Using a Model of Cervical Cancer

Sarah Dewilde, MSc, Rob Anderson, PhD

Despite early recognition of the theoretical advantages of simulations that include different population subgroups/strata and different birth cohorts, many modeling-based economic evaluations of cervical screening have been based on unrealistic single birth cohort simulations. The authors examined the effect of a multiple birth cohort simulation on the incremental cost-effectiveness estimates of cervical screening programs, compared to a conventional single cohort simulation. The choice of hypothetical cohort that starts the simulation had a major impact on the cost-effectiveness estimates: Compared with a single birth cohort simulation, the incremental cost-effectiveness of a shift from biennial to triennial screening was 30% higher when using the multiple cohort simulation. Multiple cohort simulations using the different age structures of 4 countries had little impact on the cost-effectiveness ratios (variation <5%). Future modeling-based evaluations of screening policies should better reflect the age range of the population that is targeted by carefully specifying the nature of the starting cohort(s). ***Key words:*** *cost-effectiveness; Markov model; cervical screening; cohort simulations.* ***(Med Decis Making 2004;24:486–492)***

The conventional approach to modeling the cost-effectiveness of many screening programs is to use a single birth cohort, whereby a hypothetical population—all of the same age—progresses through the defined disease states of a Markov model until some termination criterion is met. For example, evaluations of the cost-effectiveness of different cervical screening strategies have often been based on simulation of a hypothetical birth cohort of 15-, 20- or 30-year-old women.[1–7] Although such a single birth cohort approach might be appropriate for modeling screening that occurs once at a specific age, such as neonatal screening, the approach is likely to be inappropriate for assessing the costs and benefits of programs aimed at broader age ranges. Thus, in their review of modeling problems in cancer screening, van Oortmarssen and colleagues observed that "many cost-effectiveness studies consider a hypothetical birth cohort which is followed from a certain age. Few studies calculate outcomes for an actual realistic population (i.e. starting with a mixture of all ages), possibly with a history of screening."[8]

The conventional single cohort approach effectively considers the impact of a policy change only on those who are about to enter the screening age range. It does not take into account the effect of a change in screening policy on the rest of the eligible population. Moreover, if costs and effects are discounted, then the effects and costs of the program for the oldest in the target age range will be severely undervalued. Despite such observations and early recognition of this issue by the

Received 25 March 2003 from MEDTAP International® Inc., London, UK (SD), and the Centre for Health Economics Research and Evaluation, University of Technology, Sydney, Australia (RA). The first author's secondment to CHERE from York University (Department of Economic and Related Studies) was supported by CHERE through financial assistance with travel costs to and from Australia. Since submission, a version of this article has been presented at the UK Health Economics Study Group summer meeting (July 2003) at Canterbury, Kent. We gratefully acknowledge the assistance of Dr Evan Myers, of Duke University, for permission to use the cervical screening decision model and also his generous help in its adaptation to the Australian policy context. We also thank the late Dr Bernie O'Brien for comments on an earlier version of the article and 2 anonymous referees for their very useful suggestions. Revision accepted for publication 10 February 2004.

Address correspondence and reprint requests to Dr Rob Anderson, CHERE, University of Technology, Sydney, PO Box 123, Broadway, NSW 2007, Australia.

DOI: 10.1177/0272989X04268953

team that developed the MISCAN model,[9] there have been no direct comparisons using empirical data to assess how much the different cohort simulations affect cost-effectiveness estimates.

In the literature on cervical cancer screening, modeling studies that estimate costs and effectiveness by simulating birth cohorts of women of different ages are those by Sato and others,[10] those employing the MISCAN model developed by Habbema and others,[9, 11–13] and a study that reports using a starting cohort of women representative of the HIV-infected population in the United States.[14] However, Sato and others were estimating the costs and effects of "one-off" mass screening, so appropriately, they did not aggregate the estimates for different birth cohorts to generate a whole-population estimate. In contrast, Habbema and colleagues' MISCAN model is designed for micro simulation (Monte Carlo simulation), and it has been used for modeling a hypothetical population of women of different ages. However, it is not known how different the results of these studies would have been when compared to single birth cohorts simulated using the same model.

In this article, we examine the effect of a multiple birth cohort simulation on the cost-effectiveness estimates of a cervical screening program compared to estimates from a single cohort simulation. The multiple cohort simulation models the outcome of screening in a more realistic way, by reflecting the effect of screening on the whole age range affected and by taking into account the age structure of the population that would be affected by the policy decision.

METHODS

To make the comparison between single and multiple cohort, we use a 20-state Markov model of the natural progression of cervical cancer and precancerous lesions for unscreened women. The model was developed by researchers at Duke University (United States) for the US Agency for Health Care Policy and Research[7,15] and has since been adapted to re-create the processes of diagnosis, follow-up, and treatment protocols that are currently in use in the Australian health care system (CHERE, unpublished report 2002). The model incorporated age-specific transition probabilities for competing risks, such as death from other causes and benign hysterectomies. No assumptions were made about these probabilities or possible changes in human papilloma virus (HPV) incidence in different cohorts (e.g., due to changing sexual behaviors). We compared the cost-effectiveness of biennial screening, which is the current policy in Australia, with triennial screening and in both cases assumed 100% compliance with the recommended policy. All analyses were carried out using DATA 4.0 TreeAge software and Microsoft Excel.

Before running multiple simulations, the decision tree had to be adapted to allow cohorts to start at different ages. First, a single cohort simulation was run for the youngest cohort, to produce a Markov trace of all disease/health states. (This initial simulation was run using the parameter estimates for the current policy, in our case, biennial screening.) Second, the starting ages for the multiple cohort simulation must be chosen. We chose 5-year intervals (i.e., distinct cohorts starting at ages 15, 20, 25, etc.). The Markov state prevalences at these ages (from the single cohort Markov trace) are then adjusted to represent a fully alive starting cohort. These are then used as the starting prevalence variable values for each Markov state (using the DATA look-up table function).

This method of deriving initial Markov state prevalences for each cohort is admittedly unsatisfactory but unavoidable given the absence of age-specific HPV and precancerous lesion prevalence data for all Australian women (e.g., there is an early study, using data from 1970 to 1988, from only 1 Australian state; the prevalence in different cohorts was thought to be changing mainly because of evolving diagnostic criteria and because screened women were still a gradually growing minority).[16] As a partial test that the combination of initial assumptions about disease-state prevalences (in the youngest cohort) and age-related "natural history" transition probabilities (based mostly on international evidence about the natural history of disease)[15,17] produced realistic age-related prevalences, we predicted current (1998) age-specific cervical cancer incidence and age-specific cervical cancer mortality. This calibration exercise was based on a hybrid model that represented the best available evidence of the current screening behaviors in Australia (i.e., a mixed cohort of women who screened frequently, infrequently, or not at all). The calibration exercise showed good predictive accuracy for both age-specific cervical cancer incidence and mortality (CHERE, unpublished report, 2002).

In the multiple cohort simulation, we used 11 cohorts at 5-year intervals between 15 and 65 years of age, and all cohorts terminated at age 85. Figure 1 compares the survival curves of the single and the multiple cohort simulations. The starting positions of the multiple cohort survival curves represent the proportions of

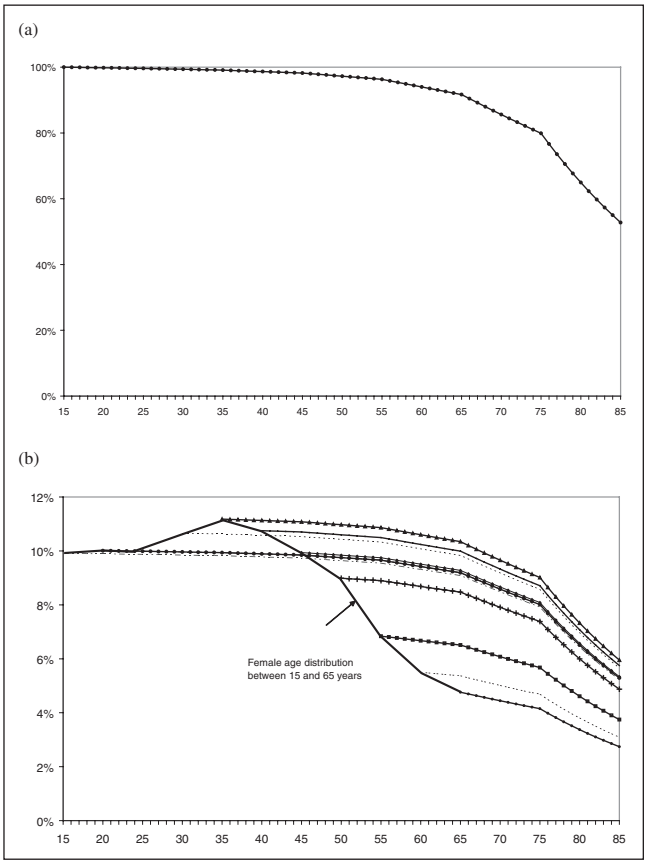

Figure 1 The survival curves of (a) the single cohort model and (b) the multiple cohort model (11 cohorts).

each age group relative to the total population of all 11 age groups (of Australian women).

Table 1 shows how the cost-effectiveness estimates are calculated and aggregated for the multiple cohort model, with biennial screening. Costs and effects (life-years saved) per person in each birth cohort are aggregated into a weighted average, according to the proportion of each age group in the target female population.

Note that these cost-effectiveness estimates were generated to demonstrate a methodological point and should not be relied on to inform actual decisions (they are generated from an earlier, slightly different version of the decision model that was used in the definitive empirical policy analysis).

RESULTS

Table 2 summarizes the cost-effectiveness results of the single and the multiple cohort simulations. A number of differences can be observed. First, both the costs and effects are lower under the multiple cohort model compared to the single cohort model. The average costs are about 12% lower under the multiple cohort model (with discounting at 5%) and about 40% lower when undiscounted. The impact of the type of model on life-years saved is similar: With the multiple cohort model, an average of 2.5 fewer life-years are saved than with the single cohort model (discounting at 5%) and 21.5 fewer life-years when undiscounted. This is unsurprising since the mean age of persons starting the multiple cohort simulation is higher than in the single cohort simulation (37 compared to 15 years), with a subsequently higher mean age during the simulation (57 compared to 47). Older age groups have fewer screening years remaining and thus accumulate lower costs and benefits of screening.

Second, the incremental cost-effectiveness ratio (ICER) from the multiple cohort simulation is about 29% higher than the ICER from the single cohort model, in both the discounted and the undiscounted case. In other words, if these data were used to consider a shift from triennial to biennial screening, the multiple cohort model would suggest that the policy shift is substantially less cost-effective than the single cohort simulation suggests. Thus, a policy maker might now be less willing to accept a policy that was originally—with a single cohort model—thought to be acceptable. This large effect of simulation type on incremental

Table 1 Calculations of the Cost-Effectiveness Estimates for the Multiple Cohort Model

Cohort	Cost (AU$)[a]	Effect (life-years)[a]	Weight	Weighted Cost (AU$)	Weighted Effect (life-years)
15+	407	19.3749	0.09774	39.79	1.894
20+	520	19.1042	0.09890	51.46	1.889
25+	517	18.7607	0.10884	56.31	2.042
30+	445	18.3348	0.10577	47.11	1.939
35+	429	17.7900	0.11128	47.80	1.980
40+	350	17.1301	0.10749	37.64	1.841
45+	322	16.2847	0.09983	32.22	1.626
50+	243	15.2949	0.09114	22.22	1.394
55+	213	14.0194	0.07004	14.93	0.982
60+	136	12.6013	0.05773	7.87	0.727
65+	83	10.7464	0.05123	4.28	0.551
Total			1.00000	361.63	16.865

a. Costs and effects are per person in the target population, discounted at 5%.

Table 2 Comparison of the Cost-Effectiveness Estimates of the 2 Screening Policies Using the Single and the Multiple Cohort Model

Intervention	Total Costs (AU$)[a]	Total Effectiveness (life-years)[a]	Incremental Cost (AU$)	Incremental Effectiveness	Incremental Cost-Effectiveness Ratio (AU$)
Discounted results (5%)					
Single cohort model					
Biennial screening	407	19.3749	—	—	
Triennial screening	295	19.3725	112.06	0.0024	47,361
Multiple cohort model					
Biennial screening	361	16.8649	—	—	
Triennial screening	253	16.8631	107.95	0.0018	61,031
Undiscounted results					
Single cohort model					
Biennial screening	1,218	63.4869	—	—	
Triennialscreening	879	63.4671	339.72	0.0198	17,100
Multiple cohort model					
Biennial screening	730	41.9795	—	—	
Triennial screening	519	41.9699	211.03	0.0096	22,061

a. Costs and effects are per person in the target population.

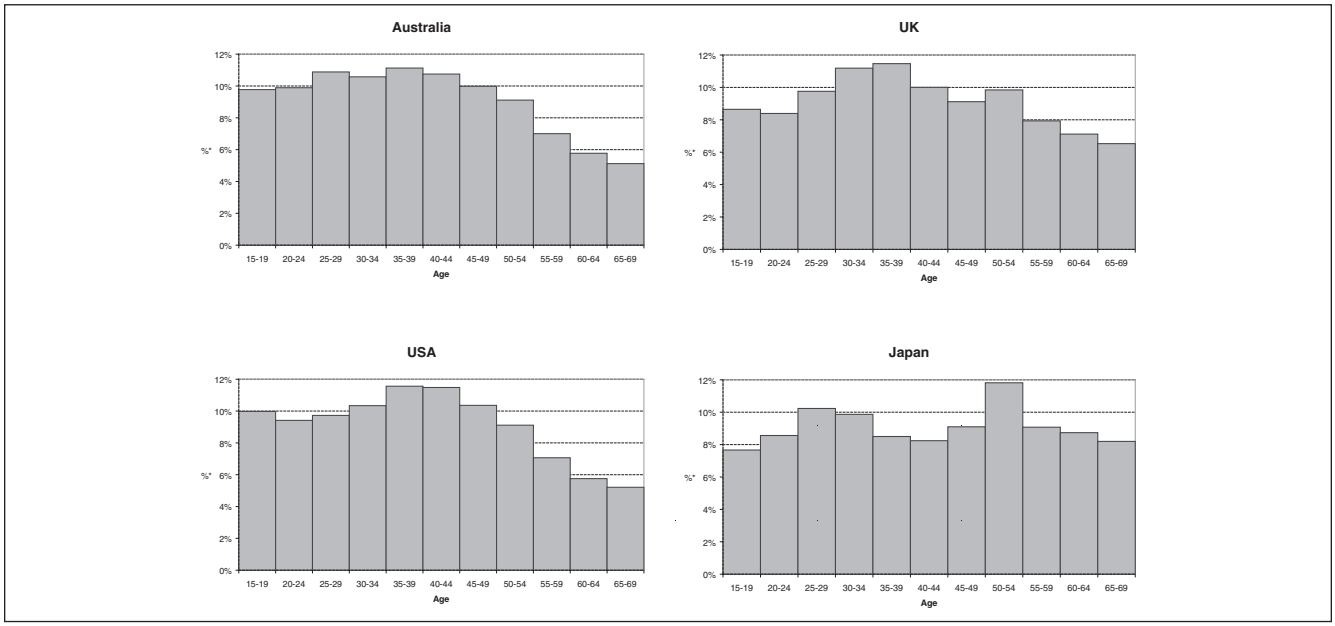

Figure 2 The female population distributions in Australia, the United Kingdom, the United States, and Japan. Data sources for female age structures are as follows. Australia: Australian Bureau of Statistics (2001). Time series spreadsheet 3201.0 (Table 9, Estimated Resident Population by Single Year of Age, Australia(a)). United Kingdom: Office for National Statistics (2001). Mid-2000 population estimates (United Kingdom, Series PE no. 3). United States: US Census Bureau (2001). Profiles of general demographic characteristics 2000 (2000 Census of Population and Housing, Summary File 1 100% data, matrices P13 and PCT12). Japan: Statistics Bureau & Statistics Centre (Japan Ministry of Public Management, Home affairs, Post and Telecommunication, 2001). Population by 5-year age group and sex, monthly estimates—total population, Japanese population, the first day, each month (October 2000–October 2001).

cost-effectiveness is evident both with and without discounting.

Next, using the multiple cohort model, we explored the influence of the target population's age distribution on cost-effectiveness. Australia has a relatively young adult population due to high levels of immigration (Figure 2). We thought that this might account for some of the difference between the cost-effectiveness of cer-

Table 3 Comparison of the Cost-Effectiveness Estimates of the 2 Screening Policies for the Single and the Multiple Cohort for the Female Age Structure of Australia, the United Kingdom, the United States, and Japan

Intervention	Total Costs (AU$)[a]	Total Effectiveness (life-years)[a]	Incremental Cost (AU$)	Incremental Effectiveness	Incremental Cost-Effectiveness Ratio (AU$)
Discounted results (5%)					
Single cohort model					
Biennial screening	407	19.3749	—	—	
Triennial screening	295	19.3725	112	0.0024	47,361
Multiple cohort model, Australia					
Biennial screening	361	16.8649	—	—	
Triennial screening	253	16.8631	108	0.0018	61,031
Multiple cohort model, United States					
Biennial screening	358	16.8316	—	—	
Triennial screening	251	16.8299	107	0.0017	61,690
Multiple cohort model, United Kingdom					
Biennial screening	348	16.6160	—	—	
Triennial screening	244	16.6143	104	0.0017	62,398
Multiple cohort model, Japan					
Biennial screening	334	16.3179	—	—	
Triennial screening	234	16.3163	100	0.0016	62,795

a. Costs and effects are per person in the target population.

vical screening using the 2 types of simulation. Therefore, we compared the Australian cost-effectiveness results with those using the starting age distribution of females in the United States, the United Kingdom, and Japan. The United Kingdom and Japan have a smaller proportion of women between 20 and 29 years than Australia does; both countries also have quite a high proportion of older (aged 40 and older) women. The age distribution of the United States resembles more closely that of Australia, except that it has fewer women in their 20s. (Note that the only change in the calculations was the age distribution of the starting cohort; the other model inputs were the same.)

Table 3 shows that Australia has the highest estimated costs and effects of screening but that, on the whole, the age distribution does not have a great influence on cost-effectiveness estimates. Even when simulating the results with a much older population than Australia's, such as Japan's, the incremental cost-effectiveness ratio changes by only 4% (from AU$61,030 to AU$62,975). So the type of simulation used influences the results more than the structure of the age distribution.

DISCUSSION

Recent guidance on good practice in decision analytical modeling focuses primarily on the structure of the model, the validation of the model estimates/inputs, and the choice between cohort or Monte Carlo simulations.[18] Although these are undoubtedly important considerations, relatively little emphasis is placed on the validity of the population that is simulated. The International Society for Pharmacoeconomics and Outcomes Research guidelines go part of the way, stating that "failure to account for heterogeneity within the modeled population can lead to errors in model results. When appropriate, modeled populations should be disaggregated according to strata that have different event probabilities, quality of life, and costs."[18] However, what we have argued and demonstrated in this article is that accounting for such heterogeneity across subjects is not just about having a valid model, structures, transition probabilities, and Markov states that accurately reflect diagnostic/prognostic and risk categories that are broadly homogeneous.[19] It is also about ensuring that the age distribution of the hypothetical population that starts in the model is similar to the one

that would be affected by the policy decision if it were made tomorrow.[20]

Our results show that it is important to think carefully about which population will be affected by screening or other health policies; modeling only a part of the relevant population can have a major impact on the cost-effectiveness estimates. This may be even more important where the age range of people affected by a policy (here, 50 years) is large relative to the likely duration of the policy chosen. Even the most carefully designed and validated model, with precise evidence-based data inputs, will produce invalid results when the model is run with an unrealistic starting population. For comparing biennial and triennial cervical screening, the incremental cost-effectiveness ratios using the 2 modeling approaches are substantially different. Despite this, in our cervical screening example, international differences in the age distribution of the female population did not greatly affect the cost-effectiveness estimates.

Enthusiasm for more "realistic" multiple cohort simulations should, however, be tempered by several further considerations. These considerations depend on the specific reasons why multiple cohort simulations might, in some contexts, be less realistic. First, without reliable age-specific data about key Markov state starting prevalences, as well as good evidence to support assumptions about possible future changes in age- or cohort-specific transition probabilities, the multiple cohort simulation may ultimately over- or underestimate the long-run costs and effects. In health policy contexts in which the probable health effects of previous screening or treatment regimes have "washed out," and where key health behaviors have stabilized, multiple cohort simulations would be less questionable. (It would, in such situations, be plausible to estimate cohort-specific state prevalences by extrapolating from current transition probabilities and prevalence rates in the youngest cohorts.)

Second, in policy contexts such as cervical screening, in which the population of those already covered by the policy is large (relative to the incoming cohorts that would be affected by it during the likely duration of the policy), the feasibility of performing a multiple cohort simulation should be explicitly explored. Conversely, in decision contexts in which the chosen policy is likely to apply for many decades (and the age range currently affected by the policy is relatively narrow), the use of a single birth cohort modeling approach is more justified. Again, however, the choice needs to be explicitly justified. Third, perhaps obviously, some policy choices (such as that between different screening frequencies or screening age ranges) are inherently time related, whereas others (such as the adoption of new sample taking or smear-reading technologies) are not. There may be arguments that for comparing non-time-related interventions, single cohort simulations give adequate estimates of longer run costs and effects.

A universal problem is that the cross-sectional data that are typically available for calibrating even single cohort models already contain cohort effects. This can be viewed as the "mirror image" to the problem of creating models that allow cohort-specific input and outcomes. Ultimately, the dilemma of choosing between single and multiple cohort simulations is therefore a particular aspect of the broader tension that pervades all modeling exercises: that between model sophistication and data availability. Modeling that is capable of informing policy makers needs to be complex enough to believably represent real-world processes (e.g., including differences between cohorts) and yet simple enough to work and—most critically perhaps—make defensible use of available epidemiological and other evidence. Cohort heterogeneity should, ideally, always be modeled, yet so much input data already contain cohort effects: The danger is that some cohort effects may, in a sense, then be double counted by their incorporation in both initial state prevalences and process probabilities.

CONCLUSION

All future modeling-based cost-effectiveness research of screening strategies should take greater account of the likely time horizon of the policy and the actual population(s) that will be affected by the particular policy decision. Critically, this means considering (and describing in the methods sections of papers) not just how long a model is run or under what conditions the simulation finishes but also the exact nature of the hypothetical cohort or individuals that start in a Markov model.

Whether the increased model complexity (and therefore the potential improvement in accuracy) of a multiple cohort model is warranted will have to be judged in relation to the corresponding availability of reliable evidence to inform the new cohort-specific model parameters. In many cases, multiple birth cohort simulations, or Monte Carlo/micro simulation of the kind possible with the MISCAN model, should be used in preference to single birth cohort simulations. Performing such multiple cohort simulations imposes extra data requirements but can be relatively straightforward with currently available software, and examples now exist for modeling a range of diseases.[21,22] How-

ever, choices about whether to adjust the model structure (to better reflect reality) or the input parameters (e.g., to negate current cohort effects, if known)—or both—will remain difficult and depend on the specifics of each analysis.

REFERENCES

1. Brown AD, Garber AM. Cost-effectiveness of three methods to enhance the sensitivity of Papanicolaou testing. JAMA. 1999;281(4): 347–53.
2. Eddy DM. Screening for cervical cancer. Ann Intern Med. 1990;113(3):214–26.
3. Fahs MC, Mandelblatt J, Schechter C, Muller C. Cost effectiveness of cervical cancer screening for the elderly. Ann Intern Med. 1992;117(6):520–7.
4. Goldie SJ, Kuhn L, Denny L, Pollack A, Wright TC. Policy analysis of cervical cancer screening strategies in low-resource settings: clinical benefits and cost-effectiveness. JAMA. 2001;285(24):3107–15.
5. Matsunaga G, Tsuji I, Sato S, Fukao A, Hisamichi S, Yajima A. Cost-effectiveness analysis of mass screening for cervical cancer in Japan. J Epidemiol. 1997;7(3):135–41.
6. Myers ER, McCrory DC, Subramanian S, et al. Setting the target for a better cervical screening test: characteristics of a cost-effective test for cervical neoplasia screening. Obstet Gynecol. 2000;96(5): 645–802.
7. Agency for Health Care Policy and Research. Evaluation of Cervical Cytology. Rockville (MD): Agency for Health Care Policy and Research; 1999.
8. van Oortmarssen GJ, Boer R, Habbema JD. Modelling issues in cancer screening. Stat Methods Med Res. 1995;4(1):33–54.
9. Habbema JD, van Oortmarssen GJ, Lubbe JT, van der Maas PJ. The MISCAN simulation program for the evaluation of screening for disease. Comput Methods Programs Biomed. 1984;20(1):79–93.
10. Sato S, Matunaga G, Tsuji I, Yajima A, Sasaki H. Determining the cost-effectiveness of mass screening for cervical cancer using common analytic models. Acta Cytol. 1999;43(6):1006–14.
11. Habbema JD, Lubbe JT, van Oortmarssen GJ, van der Maas PJ. A simulation approach to cost-effectiveness and cost-benefit calculations of screening for the early detection of disease. Eur J Operation Res. 1987;29:159–66.
12. van Ballegooijen M, van den Akker-van Marle ME, Warmerdam PG, Meijer CJ, Walboomers JM, Habbema JD. Present evidence on the value of HPV testing for cervical cancer screening: a model-based exploration of the (cost-)effectiveness. Br J Cancer. 1997;76(5):651–7.
13. Koopmanschap MA, van Oortmarssen GJ, van Agt HM, van Ballegooijen M, Habbema JD, Lubbe KT. Cervical-cancer screening: attendance and cost-effectiveness. Int J Cancer. 1990;45(3):410–5.
14. Goldie SJ, Weinstein MC, Kuntz KM, Freedberg K. The costs, clinical benefits, and cost-effectiveness of screening for cervical cancer in HIV-infected women. Ann Intern Med. 1999;130:97–107.
15. Myers ER, McCrory DC, Nanda K, Bastian L, Matchar DB. Mathematical model for the natural history of human papillomavirus infection and cervical carcinogenesis. Am J Epidemiol. 2000;151(12): 1158–71.
16. Mitchell H, Medley G. Age and time trends in the prevalence of cervical intraepithelial neoplasia on Papanicolaou smear tests, 1970-1988. Med J Aust. 1990;152(5):252–5.
17. Gustafsson L, Ponten J, Bergstrom R, Adami HO. International incidence rates of invasive cervical cancer before cytological screening. Int J Cancer. 1997;71(2):159–65.
18. Weinstein MC, O'Brien B, Hornberger J, et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practice: modeling studies. Value Health. 2003;6(1):9–17.
19. Kuntz KM, Goldie SJ. Assessing the sensitivity of decision-analytic results to unobserved markers of risk: defining the effects of heterogeneity bias. Med Decis Making. 2002;22(3):218–27.
20. Hunink MGM. Decision Making in Health and Medicine: Integrating Evidence and Values. Cambridge (UK): Cambridge University Press; 2001.
21. Hunink M, Goldman L, Tosteson A. The recent decline in mortality from coronary heart disease, 1980-1990: the effect of secular trends in risk factors and treatment. JAMA. 1997;277:535–42.
22. Loeve F, Boer R, van Oortmarssen G, van Ballegooijen M, Habbema JD. The MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. Comput Biomed Res. 1999;32: 13–33.