Statistical Methods in Medical Research 2004; **13**: 525–538

Diversity of model approaches for breast cancer screening: a review of model assumptions by The Cancer Intervention and Surveillance Network (CISNET) Breast Cancer Groups

Rob Boer Department of Public Health, Erasmus MC, Rotterdam, The Netherlands,
Sylvia Plevritis Department of Radiology, Lucas Center, Stanford, CA, USA and
Lauren Clarke Cornerstone Systems, Lynden, WA, USA

The National Cancer Institute-sponsored Cancer Intervention and Surveillance Network program on breast cancer is composed of seven research groups working largely independently to model the impact of screening and adjuvant therapy on breast cancer mortality trends in the US from 1975 to 2000. Each of the groups has chosen a different modeling methodology without purposeful attempt to be in contrast with each other. The seven groups have met biannually since November 2000 to discuss their methodology and results. This article investigates the differences in methodology. To facilitate this comparison, each of the groups submitted a description of their model into a uniformly structured web based 'model profiler'. Six of the seven models simulate a preclinical natural history that cannot be observed directly with parameters estimated from published evidence concerning screening and therapy effects. The remaining model regards published evidence on intervention effects as prior information and updates that with information from the US population in a Bayesian type analysis. In general, the differences between the models appear to be small, particularly among the models driven by natural history assumptions. However, we demonstrate that such apparently small differences can have a large impact on surveillance of population trends. We describe a systematic approach to evaluating differences in model assumptions and results, as well as differences in modeling culture underlying the differences in model structure and parameters.

1 Introduction

Early detection of cancer has a relatively long history of study by computer models.[1–10] There is currently a diversity of model approaches, some of which produce clearly different results and lead to conflicting policy recommendations. For example, model studies on colorectal cancer screening concluded to different screening tests and strategies being most efficient.[11–18] In principle, models give the opportunity to fully clarify all assumptions; however, published material generally does not include sufficient detail to make this possible and thereby limits the ability of the reader to resolve differences between model assumptions and model results. The Cancer Intervention and Surveillance Network (CISNET) Breast Cancer program is a National Cancer Institute-sponsored collaboration of seven research groups which model

Address for correspondence: Rob Boer, RAND Corporation, 1700 Main Street, PO Box 2138, Santa Monica, CA 90407-2138, USA. E-mail: boer@rand.org

 10.1191/0962280204sm381ra

the impact of screening and adjuvant treatment on trends in breast cancer incidence and mortality. This collaboration has created an unprecedented opportunity in population based cancer modeling to make a direct and detailed comparison of different models of cancer screening that are used for the same purpose.

The seven research groups that have contributed to the program's model profiler to facilitate this comparison are from the Dana-Farber Cancer Institute (F),[19–21] Erasmus MC (E),[22–26] Georgetown University (G),[27–29] University of Texas MD Anderson Cancer Center (A),[30–32] Rochester University (R),[33–35] Stanford University (S)[36,37] and the University of Wisconsin (W).[38] Each of the models will be used to simulate the trends in mammography screening, adjuvant therapy, incidence and mortality of breast cancer in the US population.

As the individual models have not been published, this article aims to present general approach among the groups and efforts that are being taken to closely compare the models and identify specific distinction between the models.

2 General approach

The general approach of most models in the CISNET breast group is remarkably similar even though they are being developed independently. Six of the seven groups model the natural history of breast cancer, with a time from birth to start of the preclinical screen detectable period, clinical diagnosis after which follows a period of survival that ends with either breast cancer death or death from other causes. The time of diagnosis can change if screening takes place during the preclinical screen detectable period, which may influence the time and cause of death. In most models, therapy influences survival from time of diagnosis (Figure 1). The models that simulate a natural history can be compared at a more detailed level.

2.1 Approaches using a natural history model

The models that are based on assumptions of the natural history can be compared in terms of the way they handle sojourn time, the mechanism of screen detection, tumor characteristics at diagnosis, survival following clinical detection and survival following screen detection.

2.1.1 Mechanism for screen detection

Early detection can occur when screening takes place during the sojourn time. In some models, screen detection is only determined by a threshold of screen detectability; in others, there is still a chance involved: sensitivity (F, G, W). From the moment of early detection, the disease can take a different course than if there was no screening – in particular, the time of breast cancer death can change.

2.1.2 Sojourn time

The models assume a sojourn time, which is the period of time before clinical diagnosis of breast cancer during which the cancer can be detected by screening. Some models (F, G) assume that this period consists of a number of disease states in which sojourn times are generated from a dwelling time distribution. Other models assume

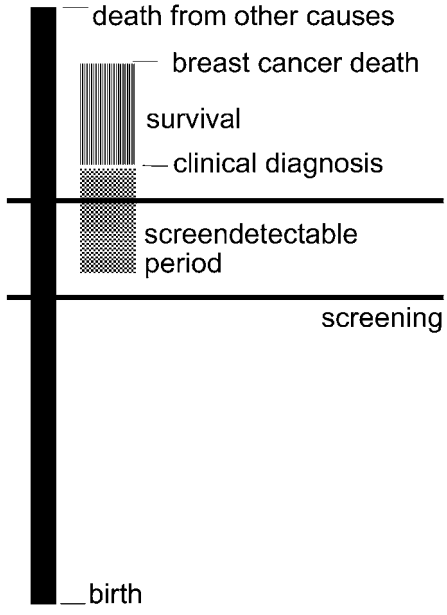

Figure 1 Generalized natural history model.

a continuous tumor growth function during the preclinical period (E, R, S, W) and use either a threshold distribution for screen detection (E, S) or sensitivity of the screening test as a function of tumor size (R, W). Owing to the differences in method of modeling and interaction with assumptions on screening sensitivity, sojourn times of the different models are not directly comparable.

2.1.3 Tumor characteristics at detection

At time of diagnosis, the cancer has attributed several characteristics that determine patient's therapy and breast cancer survival. The specific characteristics are different among the models. In all models, except E, the tumor is characterized by women's age and stage. It can also depend on calendar year (F), tumor size (E, R, S) and fatal metastasis status (E).

2.1.4 Survival following clinical detection

Most models simulated a survival distribution from time of diagnosis that is based on observed survival without screening. This survival can be improved because of application of adjuvant therapy. Model E determines the time of breast cancer death by a survival distribution from time of inception of a fatal metastasis before diagnosis, instead of from time of diagnosis.

2.1.5 Survival following screen detection

In most models, early detection often leads to detection in an earlier stage, which stage shift primarily determines an improvement of survival (F, G, R, S, W). Other models

assume that, at some time during the development of the cancer, inception of a fatal metastasis occurs that requires some period of development before the metastasis can be diagnosed, but which will lead to breast cancer death if the cancer is not diagnosed before inception of a fatal metastasis, and if mortality from other causes is not earlier (E).

Stage shift models attribute a survival distribution to screen detected cases that is different from the situation without screening, if the stage at screen detection is different. Generally, the survival distribution is estimated from stage specific survival as observed in a population based cancer registry prior to dissemination of screening. Some models apply the survival distribution from the time of screen detection, which agrees best with observed data. In order to prevent death from breast cancer during the lead time, some models apply the survival distribution of screen detected cases from the time at which diagnosis would have taken place in a situation without screening – in effect, adding the lead time (D, G), and other models revert to the original time of death of the situation without screening in individual cases, where the survival distribution from screen detection would give a time of death during the lead time (R).

2.2 Approach not based on a natural history model

Only one group (A) does not model a preclinical natural history. Instead, it simulates diagnosed cancers with their characteristics at time of diagnosis and survival, according to what is known about the dissemination of screening, tumor characteristics of screen detected and clinically detected cancer and their survival and dissemination of adjuvant therapy and its influence on survival. This model produces populations of women with breast cancer including characteristics at diagnosis and survival that vary in composition and mortality, according to uncertainty of parameter estimates as represented by prior probability distributions. If the results of a model are sufficiently close to observed mortality, then its parameter values will be included in the estimated posterior distribution. This includes a joint distribution of the contributions of the various interventions to reduction in breast cancer mortality, which informs on the relative contribution of each intervention to reduction in mortality.

3 Example of possible importance of assumptions

Generally, there appears to be a preponderance of differences that appear to be small at first consideration. However, when studying population trends that are not so pronounced, the described differences may turn out to be more influential. For example, we show the possible influence of different assumptions concerning possible mortality from breast cancer of screen detected cases during the lead time. These results are from an earlier microsimulation screening analysis (Miscan) model for breast cancer that is not used in the CISNET project, but that has been published elsewhere.[22–26] We adjusted this model only with respect to survival with and without screening. Each clinically diagnosed breast cancer is subjected to net survival distribution of stage III breast cancer in surveillance epidemiology and end results (SEER) as in Figure 2. Screen detected cancers are subjected to half the net lethality rates as clinically detected cancers (Figure 2), where the survival distribution starts either at time of screen detection

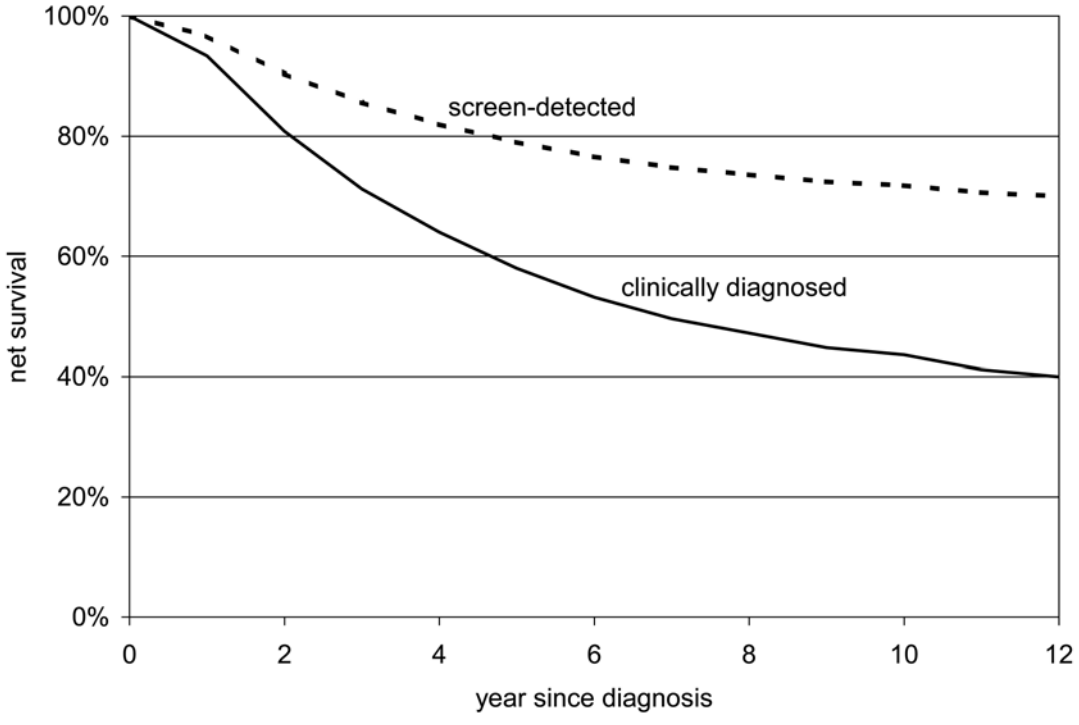

Figure 2 Tentative survival assumption.

(leading to breast cancer mortality during lead time) or at the time clinical diagnosis would have occurred without screening (no breast cancer mortality during lead time), as shown in Figure 3.

The model assumes that all women start screening from 1988 and that screening ends in 2026. As expected, incidence increases strongly during the first year of screening, is somewhat elevated relative to a model without screening during the period of repeated screening, and decreases strongly when screening stops after which incidence gradually returns to level of the situation without screening. Both model variants concerning screening effects show the same effects on incidence.

The model without mortality from breast cancer during the lead time shows a gradual decrease in mortality from breast cancer from start of screening, after which it remains at a level of around 68% of the situation without screening, and after screening stops, the mortality level gradually returns to that of the situation without screening.

The model with mortality from breast cancer during the lead time shows an increase in breast cancer mortality during the first years after start of screening. Subsequently, the mortality level gradually decreases to around 78% of the situation without screening. During the first years after screening stops, mortality decreases after which the level gradually returns to that of the situation without screening.

The specific parameter assumptions of each model will be based on several sources of empirical data. At this moment, it is not yet totally clear which data sources will

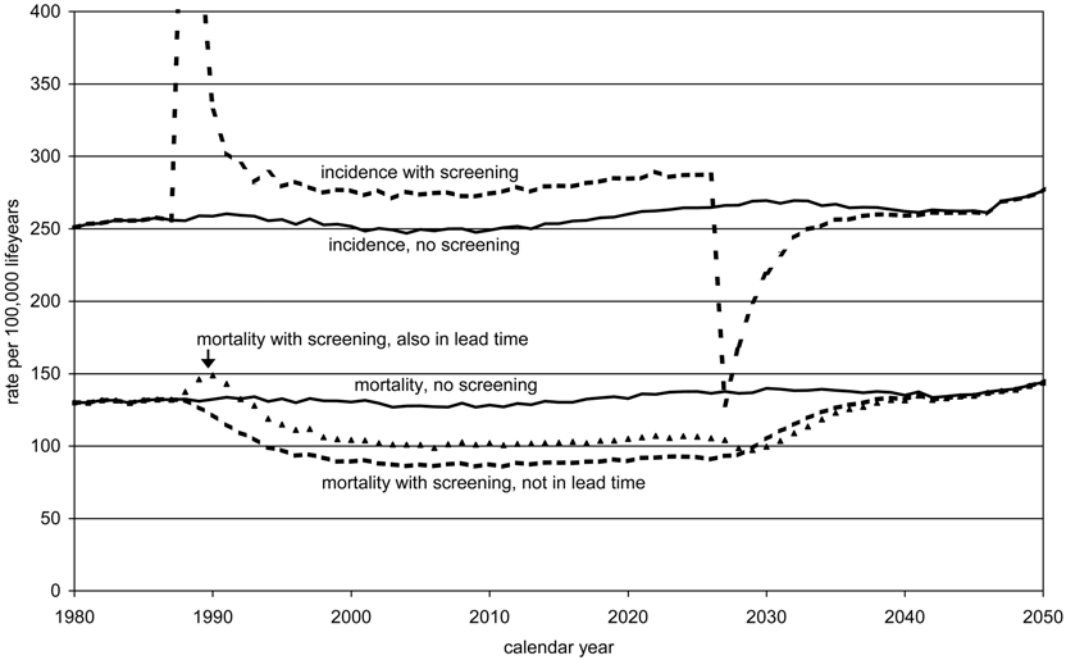

Figure 3 Modeled incidence and mortality of breast cancer with and without mortality during lead time.

be used, but they will be to some extent different for the different models. That may be another source of differences in model results. However, in general, foundation of the models on the available empirical evidence should lead to a convergence of results from different models.

Let us continue with the example of the models with and without breast cancer mortality of screen detected cases during the lead time: both models are considering a randomized trial that resulted in a mortality reduction at 15 years of follow up which is the average of what is predicted by the original two models. Each model adjusts the mortality effects of screen detection so that the model reproduces this trial result.

The predicted results will then look like Figure 4. Though, after 15 years of screening, the models will predict similar cumulative screening effects, the model with breast cancer mortality during the lead time will still predict higher mortality rates during the first years of screening, but the difference between the models reverses, and the long term predictions for cumulative breast cancer mortality prevented if the model with breast cancer mortality during the lead time were substantially higher than the other model.

4 Base case comparison

In order to facilitate an in-depth comparison of the models, the CISNET breast groups has constructed a series of model runs that asks each participating groups to model

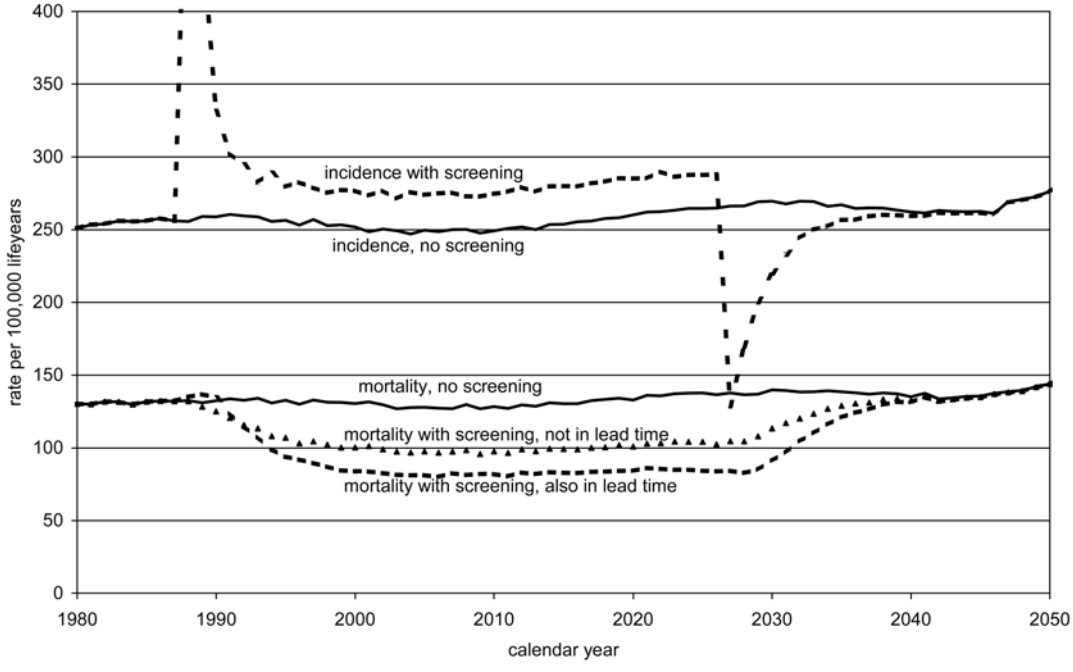

Figure 4 Modeled incidence and mortality of breast cancer with and without mortality during lead time after calibration to a trial.

a small number of scenarios concerning breast cancer in the US population using a common set of inputs ('base case inputs') and computing a common set of outputs.

The base case inputs consist of a detailed description of the population to be simulated, the mortality from other causes than breast cancer, breast cancer incidence, secular trend in breast cancer incidence and survival. It also describes detailed dissemination histories for mammography screening, and adjuvant multi-agent chemotherapy and tamoxifen.

The base case scenarios are concerned with simulating the US female population from 1975, including or excluding either or both screening and adjuvant therapy, with the goal of ascertaining the plausible ranges of contribution of each.

The models will be compared with each other and, where possible, with observed data on: breast cancer mortality by year and age; incidence by year, stage and age; mean lead time by age; overdiagnosis by age and year; detection rate at first and repeat screenings; survival by age, year since diagnosis and treatment modality; and program sensitivity.

4.1 Expected differences between model results and their interpretation among natural history models

Before comparing the models with each other, each research group is able to compare their model results with observed trends in breast cancer mortality and incidence

in the US population. Some models may be calibrated to fit the observed trends. In addition, without explicit calibration, significant differences will probably be investigated, which may lead to adjustment of the model. Convergence of modeled and observed mortality and incidence also implies convergence of the different models. Therefore, the model results concerning incidence and mortality of the scenario with screening and adjuvant therapy may turn out to be quite similar. Owing to the possible mechanisms of adjustment to observed data, it may turn out to be that model results that are not directly comparable with observed data will provide the best information about model differences.

Following is a discussion of different model results that will be compared. Most of these results do not provide immediate explanations of model differences but, with some difficulties, give a handle on further study and discussion among modeling groups.

4.2 Adjuvant therapy

The base case scenario prescribes the effects of adjuvant therapy quite closely. The main reason for possible differences in adjuvant therapy effects would be the differences in stage classification but they should lead to no more than minor differences in results. The base case output concerning survival by age, year since diagnosis and therapy will provide information on the part of a model that causes differences in therapy effects between models.

The models have much greater diversity in aspects concerning screening; therefore, the base case asks for a great amount of detail on model results that specifically reflect screening effects.

4.3 Lead time

An individual modeling group tends to think in terms of sojourn times of the preclinical phase of breast cancer or in terms of growth rates of the primary tumor. However, these cannot be compared very well between models because, for example, one model assumes that screening tests will detect the cancer with certainty when performed during the sojourn time, whereas another model assumes a probability of detection smaller than 1 and sojourn time assumptions are not directly comparable with growth rate assumptions. The actual lead times realized in a screening situation are a measure that can be compared directly between models and will show whether some models appear to assume faster developing preclinical breast cancer than others. (Lead time estimates are also age specific and can lend more information to the differences in the natural history as a function of age.)

4.4 Detection rates

Though cancer detection rates can, in principle, be observed, they are not observed in the total US population and, therefore, in this case they cannot be used to calibrate a model.

Detection rates at first screenings should be proportional to the mean sojourn time of the screen detectable phase and, therefore, provide a measure of mean sojourn time that can be compared among models. They provide less information concerning shape of the sojourn time distribution: if sojourn time distribution is exponential, then mean lead time at first screenings is equal to mean sojourn time, and if sojourn time distribution is extremely different from that, for example, constant, then mean lead time at first

screenings is still half of the mean sojourn time. In other words, detection rates at first screenings are not very sensitive for differences in shape of the sojourn time distribution.

Detection rates at repeat screenings will reflect cases of breast cancer that became screen detectable during the interval since previous screening, with the cases that were missed during the previous screening, without the prevalent cases that are currently missed and without the cases that were diagnosed during the interval. This measure is, therefore, so complicated that it will be hard to interpret on its own when trying to find causes for differences between models. Useful interpretation requires taking the context of the system into account, including detection rates at first screening and program sensitivity.

4.5 Overdiagnosis

Overdiagnosis in models, where sojourn times are simulated from the start of the preclinical phase forward, is the result of a projected time of clinical diagnosis that is later than mortality from other causes and depends largely on age of screening and the thickness of the tail of the sojourn time distribution. In models where sojourn times are simulated backward from time of actual clinical diagnosis, overdiagnosis can be simulated by assuming a separate stream of indolent cancers that are not diagnosed without screening. It is also possible to simulate overdiagnosis by continuing to generate times of clinical diagnosis after death from other causes and simulate sojourn times backward from those times.

4.6 Program sensitivity

Program sensitivity is a measure of the fraction of breast cancers that are detected by the screening program, as opposed to cancers detected during the interval between scheduled screenings. It is also a complex measure that depends on assumptions concerning sojourn time, probability of detection if screened during the preclinical phase and screening interval, and therefore difficult to interpret as cause of differences between models. The base case will compare program sensitivity considering a one-year and two-year interval since last screening, which will contribute to information concerning sojourn time distribution and test sensitivity.

4.7 Expected differences between model results and their interpretation between MD Anderson model and natural history models

On the one hand, the natural history models will use information from other sources than the US population data (e.g., from screening trials) to predict the effects of dissemination of screening and adjuvant therapy and evaluate differences between observed and predicted trends. On the other hand, the MD Anderson model (A) will adjust estimates that are initially based on sources other than the US population data, on the basis of information from the US population data. In the end, interpreting differences between expected and observed in the population, versus adjusting initially expected based on observations in the population, may not be so very different as it may seem at first. Therefore, also in this comparison, differences will probably depend to a large extent on the interpretation of the information in which initial predictions are based.

The two types of models are using the same empirical basis for estimation of effects of adjuvant therapy. However, the basis for estimation of effect of screening is quite

different: the MD Anderson model explains screening effects by observed cancers and their characteristics, particularly as associated with survival. The natural history models include a comparison with what would have happened in case a cancer was not detected by screening.

It will be difficult to analyse the reasons for possible differences between the two contrasting approaches to explaining screening effects.

5 Modeling culture

Another aspect of differences between the models does not readily appear from the previous comparison and is not always obvious in published work. However, the collaboration has demonstrated that the modeling groups can be divided into two groups. The division between the two groups sometimes resembles the rabbit–duck illusion,[39] where one just sees the rabbit and the other, the duck (Figure 5).

Though the differences between the groups are not absolute, there are a number of associated contrasts that make it difficult to communicate modeling differences. Figure 6 attempts to describe the contrasts. The contrasts are not absolute. The preference of modeling technique can tend toward using closed form mathematical formulation of the model or toward simulation, where the algorithm is too complicated to formulate in closed mathematical form.

The different model parameters can be derived from the same data or different parts of the model, including their parameters, and can be founded on different data sources. The choice of parameter values can be based on estimation or on calibration.

Though all modeling groups consider statistical validity and policy relevance to be important, when there is a tension between the two, some groups would tend to choose for statistical rigor at the expense of accessibility for policy making, and other groups

Figure 5 Rabbit–duck illusion (after Jastrow[39]).

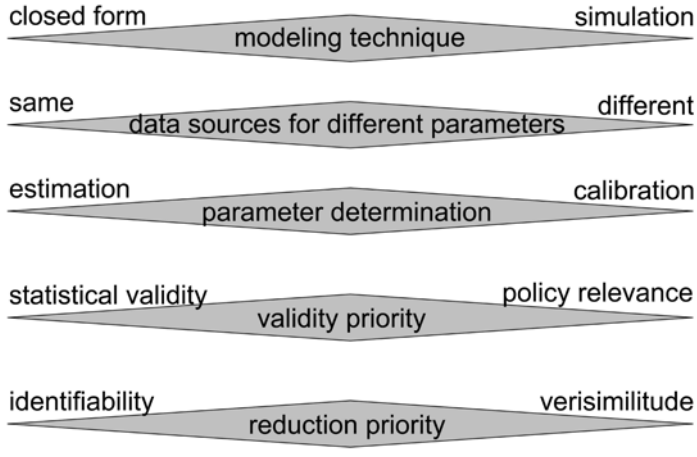

Figure 6 Contrast emphases in modeling cultures.

tend to put more emphasis on policy relevance. Similarly, some groups put more effort into resolving identifiability issues in the model, whereas others prefer maintaining higher verisimilitude.

In general, groups emphasize aspects that are all either on the left hand side of Figure 6 or on the right hand side.

6 Evaluation of model differences

The model comparison after base case results should involve an evaluation of appropriateness of model assumptions: are the assumptions made by the different models well justifiable? Currently, it does not seem likely that any model makes clearly wrong assumptions, but, in case they are found, the problem can be resolved. Remaining are different plausible sets of model structure assumptions that are used to estimate parameter values from observed data. If the different models would use all available relevant empirical information to estimate parameter values, then the differences between model results would be fully due to real uncertainty concerning structural assumptions. However, none of the models use all available information; moreover, they use different subsets of available data. It is, therefore, possible that one model shows relatively narrow confidence limits around some model results, whereas another shows relatively wide confidence limits around the model result and a different point estimate. Then the question arises: to what extent are the wider confidence limits due to using less informative data and to what extent are they due to differences in model structure?

It is not reasonably possible that the models involve all available data in their parameter estimates. For example, the Medline database includes more than 11 000

publications with Medline Substance Headings 'breast neoplasms' and 'mass screening' or 'tamoxifen'.

Using the same subset of empirical data may also present a problem. For example, some models estimate parameter values by considering individual observations and, therefore, the model estimates are limited to an empirical basis of datasets with individual data. Other models estimate from aggregate data, which may broaden the empirical basis, but reduce the informational values from a particular dataset. However, no one is using a subset of results on tamoxifen.

We expect that the model comparison will show real uncertainty in model structure assumptions and increase our understanding of why different breast cancer population surveillance models produce different results. However, it will be a considerable challenge to parse out uncertainty due to model structure and limited use of available data.

References

1 Albert A, Gertman PM, Louis TA, Liu S-I. Screening for the early detection of cancer-II. The impact of screening on the natural history of the disease. *Mathematical Biosciences* 1977; **40**: 61–109.

2 Albert A, Gertman PM, Louis TA. Screening for the early detection of cancer-1. The temporal natural history of a progressive disease state. *Mathematical Biosciences* 1978; **40**: 1–59.

3 Louis TA, Albert A, Heghinian S. Screening for the early detection of cancer. *Mathematical Biosciences* 1978; **40**: 111–44.

4 Eddy DM. A mathematical model for timing repeated medical tests. *Medical Decision Making* 1983; **3**: 45–62.

5 Knox EG. Ages and frequencies for cervical cancer screening. *British Journal of Cancer* 1976; **34**: 444–52.

6 Shwartz M. Validation and use of a mathematical model to estimate the benefits of screening younger women for breast cancer. *Cancer Detection and Prevention* 1981; **4**: 595–601.

7 Shwartz M. Estimates of lead time and length bias in a breast cancer screening program. *Cancer* 1980; **46**: 844–51.

8 Shwartz M. An analysis of the benefits of serial screening for breast cancer based upon a mathematical model of the disease. *Cancer* 1978; **41**: 1550–64.

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

10 Walter SD, Day NE. Estimation of the duration of a pre-clinical disease state using screening data. *American Journal of Epidemiology* 1983; **118**: 865–86.

11 Sonnenberg A, Delco F, Inadomi JM. Cost-effectiveness of colonoscopy in screening for colorectal cancer. *Annals of Internal Medicine* 2000; **133**: 573–84.

12 Vijan S, Hwang EW, Hofer TP, Hayward RA. Which colon cancer screening test? A comparison of costs, effectiveness, and compliance. *American Journal of Medicine* 2001; **111**: 593–601.

13 Olynyk JK, Platell CF, Collett JA. Fecal occult blood and flexible sigmoidoscopy screening for colorectal cancer: modeling the impact on colonoscopy requirements and cancer detection rates. *Journal of Gastroenterology and Hepatology* 2001; **16**: 389–92.

14 Frazier AL, Colditz GA, Fuchs CS, Kuntz KM. Cost-effectiveness of screening for colorectal cancer in the general population. *Journal of the American Medical Association* 2000; **284**: 1954–61.

15 Khandker RK, Dulski JD, Kilpatrick JB, Ellis RP, Mitchell JB, Baine WB. A decision model and cost-effectiveness analysis of colorectal cancer screening and surveillance guidelines for average-risk adults. *International Journal of Technology*

Assessment in Health Care 2000; **16**: 799–810.
16 Sonnenberg A. Cost-effectiveness in the prevention of colorectal cancer. *Gastroenterology Clinics of North America* 2002; **31**: 1069–91.
17 Bolin TD, Lapsley HM, Korman MG. Screening for colorectal cancer: what is the most cost-effective approach? *Medical Journal of Australia* 2001; **174**: 298–301.
18 McMahon PM, Bosch JL, Gleason S, Halpern EF, Lester JS, Gazelle GS. Cost-effectiveness of colorectal cancer screening. *Radiology* 2001; **219**: 44–50.
19 Zelen M, Lee SJ. Models and the early detection of disease: methodological considerations. *Cancer Treatment and Research* 2002; **113**: 1–18.
20 Lee SJ, Zelen M. Statistical models for screening: planning public health programs. *Cancer Treatment and Research* 2002; **113**: 19–36.
21 Shen Y, Zelen M. Screening sensitivity and sojourn time from breast cancer early detection clinical trials: mammograms and physical examinations. *Journal of Clinical Oncology* 2001; **19**: 3490–9.
22 van den Akker-van Marle E, de Koning H, Boer R, van der Maas P. Reduction in breast cancer mortality due to the introduction of mass screening in The Netherlands: comparison with the United Kingdom. *Journal of Medical Screening* 1999; **6**: 30–34.
23 de Koning HJ, Boer R, Warmerdam PG, Beemsterboer PM, van der Maas PJ. Quantitative interpretation of age-specific mortality reductions from the Swedish breast cancer-screening trials. *Journal of the National Cancer Institute* 1995; **87**: 1217–23.
24 van Oortmarssen GJ, Habbema JD, Lubbe JT, van der Maas PJ. A model-based analysis of the HIP project for breast cancer screening. *International Journal of Cancer* 1990; **46**: 207–13.
25 van Oortmarssen GJ, Habbema JD, van der Maas PJ, de Koning HJ, Collette HJ, Verbeek AL, Geerts AT, Lubbe KT. A model for breast cancer screening. *Cancer* 1990; **66**: 1601–12.
26 Habbema JD, van Oortmarssen GJ, Lubbe JT, van der Maas PJ. The MISCAN simulation program for the evaluation of screening for disease. *Computer Methods and Programs in Biomedicine* 1985; **20**(1): 79–93.
27 Mandelblatt JS, Lawrence WF, Womack SM, Jacbson D, Yi B, Hwang YT, Gold K, Barter J, Shah K. Benefits and costs of using HPV testing to screen for cervical cancer. *Journal of the American Medical Association* 2002; **287**: 2372–81.
28 Mandelblatt JS, Lawrence WF, Gaffikin L, Limpahapm KK, Lumbiganon P, Warakamin S, King J, Yi B, Ringers P, Blumenthal PD. Costs and benefits of different strategies to screen for cervical cancer in less-developed countries. *Journal of the National Cancer Institute* 2002; **94**: 1469–83.
29 Schechter CB. Cost-effectiveness of rescreening conventionally prepared cervical smears by PAPNET testing. *Acta Cytologica* 1996; **40**: 1272–82.
30 Berry DA. Screening mammography for women in their forties. *Breast Disease* 1998; **10**: 23–32.
31 Berry DA. Benefits and risks of screening mammography for women in their forties: a statistical appraisal. *Journal of the National Cancer Institute* 1998; **90**: 1431–39.
32 Gordis L, Berry D, Chu S, Fajardo L, Hoel D, Laufman L, Rufenbarger C, Scott J, Sullivan DC, Wasson J, Westhaff C, Zern R. National Institutes of Health Consensus Development Conference Statement: Breast Cancer Screening for Women Ages 40–49, January 21–23, 1997. National institutes of health consensus development panel. *Journal of the National Cancer Institute* 1997; **89**: 1015–26.
33 Yakovlev AY, Hanin LG, Rachev ST, Tsodikov AD. A distribution of tumor size at detection and its limiting form. *Proceedings of the National Academy of Sciences of the United States of America* 1996; **93**: 6671–75.
34 Tsodikov AD, Asselain B, Yakovlev AY. A distribution of tumor size at detection: an application to breast cancer data. *Biometrics* 1997; **53**: 1495–502.
35 Yakovlev AY, Tsodikov AD, Boucher K, Kerber R. The shape of the hazard function in breast carcinoma: curability of the disease revisited. *Cancer* 1999; **85**: 1789–98.
36 Plevritis SK. A mathematical algorithm that computes breast cancer sizes and doubling times detected by screening. *Mathematical Biosciences* 2001; **171**: 155–78.
37 Plevritis SK. A framework for evaluating the cost-effectiveness of MRI screening for

breast cancer. *European Radiology* 2000; **10**: S430–32.
38 Craig BA, Fryback DG, Klein R, Klein BE. A Bayesian approach to modelling the natural history of a chronic condition from observations with intervention. *Statistics in Medicine* 1999; **18**: 1355–71.
39 Jastrow J. *Fact and fable in psychology*. Boston: Houghton, Mifflin, 1900.