Chapter 13:

A Comparative Review of CISNET Breast Models Used To Analyze U.S. Breast Cancer Incidence and Mortality Trends

Lauren D. Clarke, Sylvia K. Plevritis, Rob Boer, Kathleen A. Cronin, Eric J. Feuer

The CISNET Breast Cancer program is a National Cancer Institute–sponsored collaboration composed of seven research groups that have modeled the impact of screening and adjuvant treatment on trends in breast cancer incidence and mortality over the period 1975–2000 (base case). This collaboration created a unique opportunity to make direct comparison of results from different models of population-based cancer screening produced in response to the same question. Comparing results in all but the most cursory way necessitates comparison of the models themselves. Previous chapters have discussed the models individual in detail. This chapter will aid the reader in understanding key areas of difference between the models. A focused analysis of differences and similarities between the models is presented with special attention paid to areas deemed most likely to contribute substantially to the results of the target analysis. [J Natl Cancer Inst Monogr 2006;36:96–105]

The CISNET Breast Cancer program is a National Cancer Institute–sponsored collaboration composed of seven research groups that have modeled the impact of screening and adjuvant treatment on trends in breast cancer incidence and mortality over the period 1975–2000. The group modeled the target population under several scenarios, some hypothetical. These scenarios were as follows: no screening and no adjuvant therapy, screening only, chemotherapy only, tamoxifen only, adjuvant treatment only (no screening), and screening and adjuvant treatment. This collaboration created a unique opportunity to make direct comparison of results from different models of population-based cancer screening. Comparing model results inevitably leads to comparing models themselves. Because the models in the CISNET collaborative were developed independently, there are many differences as well as similarities in the implicit and explicit building blocks inherent in each model. This paper is a focused analysis on aspects of the models that are most important to the results examined in the primary, (base case) analysis, which quantifies the impact of adjuvant therapy and screening mammography on breast cancer mortality *(1)*, and on aspects that are deemed most informative in describing the diversity of approaches used. Challenges in documenting and comparing disparate modeling efforts and the tools used to mitigate these challenges in the CISNET program are discussed briefly.

Methods

As each of these models was applied to the base case question and respective results were compared, it became necessary to develop tools to aid in the higher-level metacomparison of the models themselves. To facilitate development of comparative model documentation, a template-based system was devised to capture the salient parts of each model.

To manage the model documentation process and facilitate comparison of those documents, an Internet-based documentation software tool was developed to support the CISNET program modelers. This Model Profiling Framework provides both templated and free-form document types for the description of model components. The process of developing models of real-world phenomenon is, by nature, iterative (see “Discussion”).

Documentation of these models is also iterative as new understanding requires modification of both the model and the documents used to describe it. Also, it is difficult at the outset to determine where the most detailed documentation will be needed to determine the root of key differences in model results. To this end, the Model Profiling Framework was designed to allow modification of documentation, introduction of new templates, and even modification of existing templates as needed.

Each group participating in this collaboration provided documentation for their models in a series of subdocuments based on common templates. These subdocuments are designed to be largely independent from each other and naturally fall into a detail-based hierarchy starting at a general level and becoming more specific. The Model Profile is defined as the set of these subdocuments. The various subdocuments answer a specific set of questions outlined in the template for each particular document. The Model Profile, the collection of these documents, is designed to be read in a nonlinear fashion according to the priorities of the reader. When gathered together, the Profiles from various groups may be compared at the subdocument level either via a Web-based interface or via a static PDF document. Individual and joint model profiles are available for download *(2)*.

Affiliations of authors: Cornerstone Systems, Lynden, WA (LDC); Department of Radiology, Stanford University, Stanford, CA (SKP); Rand Corporation, Santa Monica, CA (RB); Statistical Research and Application Branch, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD (KAC, EJF).

Correspondence to: Lauren D. Clarke, MS, Cornerstone Systems Northwest Inc., 8665 Berthusen Rd., Lynden, WA 98264 (e-mail: lauren@cornerstonenw.com).

DOI: 10.1093/jncimonographs/lgj013

Table 1. Overview of primary purpose

Model*	Purpose	Ref
E	To analyze and explain results of cancer screening trials, to predict and compare the (cost-) effectiveness of different screening policies, and to monitor the results of population screening programs	*(3)*
S	To quantify the impact of screening mammography and adjuvant therapy on breast cancer mortality trends from 1975 to 2000 and evaluate the impact of new screening tests and treatment strategies on future trends	*(4)*
M	To provide estimates (and their associated uncertainties) of the relative contributions of screening mammography, tamoxifen use, and improvements in chemotherapy to the observed decrease in U.S. breast cancer mortality since 1990	*(5)*
D	To predict national mortality trends, predict the outcome of early detection clinical trials, evaluate service programs, and investigate different screening schedules to compare mortality benefit	*(6)*
G	To examine the benefits (and costs) of cancer control interventions in differing age and race/ethnic groups	*(7)*
R	To design explanatory and predictive tools for quantitative description of the effects of breast cancer screening for various screening strategies	*(8)*
W	To generate a realistic virtual Wisconsin cancer registry of incident breast cancers for women residing in Wisconsin from 1975 to 2000 and to simultaneously replicate age-specific breast cancer mortality in this population during the same period; also to explore ramifications of alternative programs of screening and treatment for breast cancer	*(9)*

*Model group abbreviations: D (Dana-Farber Cancer Center), E (Erasmus University Rotterdam), G (Georgetown University Medical Center), M (University of Texas M. D. Anderson Cancer Center), S (Stanford University), R (University of Rochester), W (University of Wisconsin–Madison).

Basic Model Differences and Similarities

Before comparing specific model components, we identify at high-level differences of the environment in which these models were developed. As with any scientific endeavor, the primary aims and guiding principles are critically important factors in determining the research course. We focus on particular model details in a later section. Modeling group abbreviations used in this paper are provided in Table 1.

Though independently developed, the CISNET breast models shared some common features. Six of the seven groups model the natural history of breast cancer. Typically, the simulated natural history includes a time from birth to start of the preclinical screen-detectable period followed by clinical diagnosis, after which follows survival that ends with either breast cancer death or death from other causes. If applied during the preclinical phase, screening could influence the timing and characteristics of disease at diagnosis (usually via a shift in stage). This in turn could affect survival so that the patient lives past the point of death under clinical diagnosis and can influence the cause of death if cancer-specific survival under screening is extended past the age at other-cause death. In most models, treatment influences survival from time of diagnosis, albeit in different ways. Despite these broad similarities, the underlying assumptions and data used to inform the models varied considerably between models in ways that can affect the sometimes subtle population trends under analysis.

Model Building Philosophies

This collaboration are used multipurpose models. That is, the authors of the models envisioned purposes for their model that lay outside the particular question posed by the base case.

For CISNET collaboration, these models were adapted to answer the common "base case" question: "What is the Impact of Adjuvant Therapy and Screening Mammography on U.S. Breast Cancer Mortality: 1975–2000?" Which is to say, the models were capable of considering the target population under several scenarios, some hypothetical. These scenarios are as follows: no screening and no adjuvant therapy, screening only, chemotherapy only, tamoxifen only, adjuvant treatment only (no screening), and screening and adjuvant treatment.

Although the models were capable of addressing the CISNET base case question, there existed a good deal of heterogeneity in the primary purpose of the various models, which in turn leads to differences in modeling approaches. Table 1 provides summaries of the base purposes for which the various models were built as well as a reference to the chapter in this monograph where the corresponding model is described in detail. Table 2 outlines some key high-level differences between the models. Although the guiding philosophies are listed, these philosophies were often shared to some degree among all the models. To model is to choose, and each group started with a certain set of values that in aggregate amounted to model philosophies, which in turn influenced the modeling choices. These core values were driven largely by the target application(s) of the model but also by available data, epidemiological principles, assumptions held, pragmatism during model building, and future plans for the model. There is a certain amount of pragmatic overlap that occurs during the model building process that makes it impossible to assign a blanket philosophy to any particular modeling effort to the exclusion of all others; however, knowing the general philosophical approach taken in each effort can aid in understanding model design. Models designated as having a "Comprehensive" primary philosophy in Table 2 took a systems engineering approach and attempted to capture at least some of the underlying, physically meaningful components of the system being modeled. In these models, the model inputs are interpretable with regard to the phenomenon being studied, e.g., exam sensitivity, stage shift, and sojourn time. Sometimes these parameters are not (ethically) observable (e.g., thresholds for tumor detection, tumor growth rates) but are estimated via a calibration step where model

Table 2. Overview of characteristics

Model*	Guiding building philosophy	Type
E	Comprehensive	Simulation
S	Comprehensive	Hybrid†
M	Observation focused	Simulation
D	Comprehensive	Analytic
G	Observation focused	Simulation
R	Observation focused/comprehensive	Hybrid†
W	Comprehensive	Simulation

*Model group abbreviations: D (Dana-Farber Cancer Center), E (Erasmus University Rotterdam), G (Georgetown University Medical Center), M (University of Texas M. D. Anderson Cancer Center), S (Stanford University), R (University of Rochester), W (University of Wisconsin–Madison).

†Hybrid models use both analytic and simulation components.

outputs are compared against empirical data such as clinical incidence and unknown parameters are adjusted (within certain constraints) such that the model reproduces observed results. Those labeled as "Observation Focused" typically trade comprehensiveness for a more data-driven approach that incorporates fewer assumptions about underlying mechanisms of tumor growth or how screening works. The "observation focused/comprehensive" philosophy seeks to maximize the flexibility and complexity of the model under the constraint of its identifiably. This approach represents a further restriction on modeling unobservables in that calibration is applied to parameters only if there are biological grounds to believe that the parameter in question may differ between the data used to build the model and the situation being simulated. For example, risk factors and sensitivity of screening procedures may conceivably vary between countries/datasets and may be calibrated, whereas growth rates would be considered a more fundamental biological parameter that should not be calibrated.

Again, the generalized guiding philosophies were not mutually exclusive in practice.

Techniques and Philosophies in Calibration and Validation

Almost by definition, model building involves combining empirical evidence with assumptions about underlying processes. Assumptions may be necessary for simplification or be based on processes that are not directly observable. During the model development process, empirical data can play several different roles. First, model parameters might be estimated from these data. Second, model structures that support less well-understood aspects of the model's domain might be modified on the basis of empirical evidence. Third, deeper unobservable or unknown model parameters might be estimated by comparing model results against data not already used in the model building process. Finally, the model's predictive ability might be tested by having it simulate a situation for which previously unused empirical data are available. These varied uses of data during the model building process necessitated a set of working definitions as follows:

Estimation. Parameter estimation is the process whereby parameters that inform the larger model are estimated from observed data and various analytical techniques to produce a result that will be consumed by the model as an input. Parameter estimation is done *independently* of the larger model under construction, although basic statistical model assumptions are used in the estimation process. A related activity, calibration (see below), is not independent of the model under construction.

Calibration. Calibration is the process whereby unknown parameters are estimated using the model itself (versus some other form of estimation outside the model). Typically this is done via minimization of the residuals between model results and empirical data collected in a situation the model is set up to replicate.

Verification. The verification process is a form of software acceptance testing in which it is determined whether the implemented model is a faithful representation of the conceptual model. This process answers the question: "Does the model respond as expected in simplistic boundary conditions such as zero incidence, no other-cause mortality, no benefit from screening or treatment, and other (typically contrived) sets of input parameters where expected model results may be stipulated before results are generated?"

Validation. Validation is an evaluation of the model performed by measurement of the model's ability (without further adjustment) to replicate observed results in a given scenario. Some groups further refine this activity by defining two variants: *internal* and *external* validation. Internal validation is similar to verification in that it tests the model's ability to reproduce datasets used (at least in part) in its construction, perhaps using a portion of the building datasets and testing the ability of the model to predict the remainder. External validation tests the model's ability to reproduce observed results not used in its construction and in general. Various criteria may be set up to accept or reject a model on the basis of its ability to validate. In short, validation challenges the model assumptions, particularly those assumptions that are most important for future extrapolations where observed data is not available.

Data used in model verification are typically contrived to uncover subtle problems with the model and to minimize the effort needed to predict correct responses from the model. For example, the model may be run with no screening, or no death from cancer or no treatment. Results from these strictly hypothetical scenarios are relatively easy to predict separate from the model and so the model's basic integrity can be assessed without much effort in these extreme cases. Datasets used in calibration and validation come from varied sources: results from randomized controlled trials, autopsy studies, observational studies, population based surveillance prior to intervention, and population-based incidence and mortality trends. There is almost always a shortage of data suitable for use in calibrating or validating a model. Collecting such data is both expensive and time consuming and modeling efforts must proceed, almost by definition, in the face of imperfection in the underlying data. As a result, careful decisions must be made to ensure that adequate, independent data remain for validation after calibration is complete. Philosophies on these issues varied between the groups. Results from specific validations for the models may be found in the group-specific chapters in this monograph *(3–9)*.

The goals of the breast cancer base case analysis generated some interesting philosophical debates on the issue of appropriate validation and calibration targets. The aim of the base case analysis is to *partition* the impact of screening and therapy on observed population-based mortality trends. The mortality itself is not the primary target, but rather the apportioning of the mortality trends between two attributable causes. As such, it may be reasonable to use overall mortality as a target during the calibration phase of model construction, thus assuring that observed mortality would be reproduced by the model a priori. However, given the complexity of the underlying factors influencing mortality trends, it may be unreasonable for any model to predict observed mortality exactly. Indeed, given factors such as risk factor dynamics in the population, changes in prevention, changes in diagnostic technologies, and coding changes, one should not expect a model to reproduce observed population mortality exactly without accounting for these additional factors. These problems may be mitigated by the addition of an "other factors" term in the calibration process, which would represent these factors not modeled explicitly. However, care must be taken here as well since overall mortality is closely associated with the outcomes of interest, namely, the contributions of screening and therapy to said mortality trends. As such, there is a risk of one causal agent (screening or treatment or "other") being arbitrarily favored over another during the calibration phase. Also, since mortality is an outcome at the extreme end of the cancer screening modeling process, calibration to it may amount to a confounding interplay between misspecification (misrepresented or missing model components) and unidentifiability (multiple parameter sets fit equally well) in the model, and care must be taken to limit the number of parameters and components under consideration in any calibration process and to adequately explore all calibrated parameter values that fit equally well. Given these factors and the various philosophies described earlier, it is not surprising that the various modeling groups split into several categories with respect to calibration.

The modeling efforts in the CISNET breast program may be roughly partitioned into three philosophies of calibration termed here as “high,” “medium,” and “low,” where the terms indicate the propensity for the modeling effort to adjust model inputs to match observed overall mortality. Model M used a high degree of calibration to the overall mortality trends, whereas W used high-degree calibration to incidence trends including breast cancer in situ. In both efforts, prior parameter distributions are defined by available data or expert opinion and the model is run with a parameter set chosen at random from the parameter space defined by the prior parameter distributions. Parameter sets are accepted or rejected on the basis of their conformity to observed mortality (M) or incidence (W) trends. Thus, through a process of Bayesian updating, posterior distributions of parameters are determined. Provided the prior distributions of parameters and the internal structures of these models can replicate the observed results, this process will, by design, produce results that closely match the observed mortality trends. Here the posterior parameter distributions are the product of interest, not the goodness of fit to observed mortality.

The R model is categorized as “medium” with respect to calibration. In this effort, observed mortality trends were used as calibration targets to perform small adjustments to unknown parameters. In particular, Canadian screening trial data were used to build the analytic model and calibration to distribution of tumor size at diagnosis was used to adjust mammography operating characteristics to account for differences in practice between Canada and the United States. Calibration was also used to adjust scaling and change-point parameters for treatment effects.

The other models, S, G, E, and D, take a “low” approach to mortality calibration in that they calibrate to endpoints other than mortality (S, G, E)—or they do no calibration at all (D).

It is important to keep these different calibration approaches in mind when looking at results from these models *(10,11)* as they will have an impact on the overall fit to observed various outcomes. It is not clear how these approaches to calibration to the mortality endpoint may influence the base case analysis results, namely, the contribution of screening and treatment to mortality trends.

Model Inputs

While not exactly components of the models themselves, inputs in a given model can have a strong influence on the model structure, capabilities and results.

Common available input data. Several common inputs were developed and made available for use in performing the base case analysis. These inputs were developed in collaboration with NCI staff and the various modeling teams in CISNET. These inputs represent the raw materials for the base case analysis and are based on empirical data collected from various sources. Chapters 2–5 in this monograph describe these inputs in detail *(12–15)*. Not all models used all the available inputs, and approaches to input usage varied across the different modeling teams (Table 3). For example, some groups used some of the common inputs as targets for calibration processes in which deeper, unobservable model parameters were estimated. In this way, the base case inputs were used to derive other related model parameters so that the model would replicate the provided base case input.

Table 3. Base case input usage by model*

	Model						
Input	D	E	G	M	S	R	W
Dissemination of adjuvant therapy	U	U	U	U	U	N	U
Dissemination of mammography	E	U	U	U	U	U	U
Other-cause mortality	N	U	U	U	U	N	U
Secular trends in underlying risk of BC incidence	U	C	U	N	U	C	C
Prescreening estimates of BC survival	U	E	U	U	U	N	A
Prescreening stage distribution at diagnosis	A	E	U	A	E	N	C
Adjuvant treatment effectiveness	U	C	U	A	U	N	E
Prescreen BC mortality	N	A	N	N	N	C	U
Observed SEER incidence rates 1975–1999	E	E	C	U	C	E	C

*N = Not used; U = uses in form provided to all groups (standard CISNET input); C = calibrates to the CISNET input (it is an output, and other inputs are calibrated to reproduce the provided input); E = standard CISNET input is used to estimate model parameters; A = alternative data used; BC = breast cancer; SEER = Surveillance, Epidemiology, and End Results.

Additional inputs. In addition to these base case inputs, several groups used alternative inputs to accommodate certain parameter estimation needs. Group E used the Swedish Two County breast cancer screening trial data to estimate certain deep model parameters needed for the fatal diameter aspect of that model *(18)*. Also, the R group used data from the Canadian National Breast Screening Studies *(20)* to estimate tumor growth rates, screen sensitivity, and other cohort-specific parameters, sometimes adjusting for differences in population. Please refer to the individual model description chapters in this monograph for a complete list of inputs used in each case.

Model Details

As noted, the philosophical approaches to model building varied among the participating groups. However, the nature of the common base case analysis dictated, to some extent, that a certain set of core components be present in each of the models. We now turn our attention to these more detailed aspects of the modeling process.

Population Modeling Issues

One of the challenges inherent in modeling a dynamic population over a fixed period is that simulated time may in fact be considerably longer than the period of interest. Although the target

Table 4. Approaches to calibration to overall mortality*

Model	Calibration level	Inputs calibrated	Calibration dataset
D	Low (none)†	—	—
G	Low	Sojourn time, operating characteristics of screening	SEER incidence, stage distribution
E	Low	Tumor diameter at diagnosis, Fatal diameter parameters	1975 SEER incidence, Two County Study, 1975 SEER survival, historical survival treatment effect
S	Low	Mean growth rate, median detection threshold	SEER incidence
R	Medium	Natural history, operating characteristics of screening, scale and change point for treatment effects	SEER incidence SEER mortality
M	High	All priors	SEER mortality Screening and treatment effects
W	High	Natural history, operating characteristics of screening	SEER incidence SEER mortality

*Model group abbreviations: D (Dana-Farber Cancer Center), E (Erasmus University Rotterdam), G (Georgetown University Medical Center), M (University of Texas M. D. Anderson Cancer Center), S (Stanford University), R (University of Rochester), W (University of Wisconsin–Madison). SEER = Surveillance, Epidemiology, and End Results; — = none.

†The D group did shift their overall mortality (post simulation) result to match 1975 U.S. mortality, which has no effect on estimates of mortality reduction from 1975.

period of observation for the base case analysis is 1975–2000, in many modeling approaches it is necessary to generate simulated events prior to and after this window of time. For models that simulate individuals from birth, it may be necessary to simulate of birth cohorts from about 1890 (85-year-olds in 1975) to 1970 (30-year-olds in 2000). Also, for most microsimulation approaches, cancer incidence must be generated both before and after the period of interest. For example, simulation results prior to 1975 must be generated to obtain the correct prevalence in 1975. And for models that use approaches to natural history that start with clinical incidence and work backward to onset (G, S), simulation must continue after 2000 to generate screen-detectable, preclinical disease in the last years of the interval and thus avoid a lead time–driven tapering of incidence in the years near the end of the period of interest. These differences in approach to population generation have a sizable impact on implementation and performance issues; however, due to age adjustment in the final results it is unlikely that they had substantial impact on the results.

Incidence

For incidence, the models may be partitioned into two sets. Those that calibrated to some form of observed incidence, E, S, G, R, and W, and those that did not, M and D. Those that did calibrate to incidence may be further divided into those that calibrated Surveillance, Epidemiology, and End Results (SEER) incidence over the years 1975–2000 (S, G, R, and W) and those that used alternative incidence measures (E). See Table 4 for a breakdown of these differences. In general, incidence calibrators did their calibrations to estimate parameters driving unobservable aspects of their natural history models.

Natural History and Survival

Of the seven models in the CISNET breast program, all but one (M) incorporated a natural history component in which transitions through disease states were generated and tracked on the basis of assumptions, data, and expert opinion surrounding the natural progression of breast cancer. The models that are based on assumptions of the natural history can be compared in the way that they handle sojourn time, the mechanism of screen detection, tumor characteristics at diagnosis, survival following clinical detection, and survival after screen detection.

Sojourn time of preclinical disease. The models assume a sojourn time, which is the time before clinical diagnosis of breast cancer during which the cancer can be detected by screening. Some models (D and G) assume that this period consists of several discrete disease states from which sojourn times are generated as an exponential dwelling time distribution. Other models assume a continuous tumor growth function during the preclinical period (E, R, S, and W) and use either a threshold distribution for screen-detection (E and S) or sensitivity of the screening test is a function of tumor size (R and W). Due to the differences in methods of modeling and the interaction with assumptions on screening sensitivity including detection thresholds, sojourn times of the different models are not directly comparable. See Table 3 of chapter 14 for a detailed list of approaches to preclinical duration *(11)*.

Mechanism of screen detection. Early detection can only occur when screening takes place during the preclinical sojourn time. In some models screen detection is determined solely by a threshold (usually size based) of screen detectability (E, S), whereas in others there is still an element of chance involved (D, G, W, R) at the time of the screening test. From the moment of early detection, the disease can take a different course from that taken if there were no screening; in particular, the time of breast cancer death can change, and if cancer death occurs at a time after the predetermined time of other-cause death, the cause of death can also change, thus affecting disease mortality rates in the simulated population. See Table 2 in chapter 14 *(11)*.

Tumor characteristics at detection. At time of diagnosis the cancer has several characteristics that determine the patient's subsequent therapy and breast cancer survival. The specific characteristics are different among the models. In all models, except E, the tumor is characterized by the women's age and stage. Survival can also depend on calendar year (D), tumor size (E, R, S), and discrete fatal diameter status (E). Four of the models (G, E, W, M) included ductal carcinoma in situ (DCIS) disease, while avoiding modeling this phenomenon explicitly by incorporating DCIS with very localized disease with very good prognosis. On balance, it seems these approaches made little consistent impact on mortality [see discussion in chapter 15 *(10)*]. Handling of in situ disease will, however, probably have a much larger impact in cost-effectiveness analyses performed with these models.

Only one group (M) does not model a preclinical natural history of any kind. Instead, it simulated 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 patients 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 parameters' estimates as represented by prior probability distributions. As described earlier, if the model's results under a particular set of parameters drawn from the priors are sufficiently close to observed mortality, then those parameter values will be included in the estimated posterior distribution. These generated posterior distributions include a joint distribution of the contributions of the various interventions to reduction in breast cancer mortality, which informs the primary result of the base case analysis: the relative contribution of each intervention to reduction in mortality.

Screening Dissemination

Except group D, all the groups used the same screening dissemination generation process *(15,23)*. Dana-Farber used the parameters of the screening dissemination generation program rather than the output from it. This group partitioned screeners into three types of screening intervals (1 year, 2 year, and 5 year), whereas the screening dissemination program used by the other groups would introduce variability in the screening patterns within women (with an affinity to a general screening behavior or annual, biennial, and irregular). On balance, these differences in approach caused the Dana-Farber model to generate shorter times between subsequent screening exams and thus simulate more screening. This higher rate of screening, combined with Breast Cancer Surveillance Consortium–based screened stage distributions that were unique to this model, explain much of the overall higher benefit associated with screening reported by the Dana-Farber group.

Survival Benefit Attribution

Survival after clinical detection. Most models simulated a survival distribution from time of diagnosis that is based on observed survival without screening, i.e., survival in the 1970s. 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 the time the tumor reaches a fatal diameter threshold (Weibull distributed) before diagnosis, instead of from time of diagnosis. Thus, here screen detection has no effect unless it occurs prior to the time the tumor reaches the fatal diameter. Model W assigns an age of death for noncured patients after progression to late-stage disease by using SEER survival curves from clinically detected late-stage disease in the prescreening era.

Survival after screen detection. Early detection can lead to detection in an earlier stage, which according to the basic assumptions of many of the models, will lead to better prognosis and possibly longer disease-specific survival. Generating a breast cancer survival curve for a screen-detected case is challenging because it is expected that in addition to a potential survival benefit, there is a lead time and length bias component in survival when disease is screen detected. The approach taken to deal with these issues varies depending on model structure and modeling philosophy. This aspect of the models is of critical importance and comparison is often confounded by different approaches and subtle definition differences inherent in independent model building. For example, some groups model individual life histories in which the simulated individuals are placed both in intervention and control arms and events in either case are tracked and compared. Thus a perfectly matched experiment is simulated and individual lead times are known. Other models simulate population strata in which a distribution of lead times may be determined, whereas individual lead times cannot. In these two cases, approaches to dealing with lead time bias in survival reporting can be different because of what is known in each case.

Length biased sampling (LBS) is a statistical sampling phenomenon that arises in screening wherein disease with longer preclinical periods tend to be detected more readily than those with shorter sojourn times. This is a naturally occurring sampling phenomenon that may contribute (in part) to the stage shift seen in screen-detected cases. The implications of LBS on screening outcomes are not well known. One possible hypothesis is that LBS causes screening to detect more indolent disease that, even if clinically detected, would not contribute to mortality *(24)*. If this were true, one would expect interval cancers to have, on balance, worse prognosis and higher mortality, and screen-detected cancers would enjoy better survival and mortality even when adjusting for lead time and controlling for age and stage. Although all the CISNET models simulate LBS, they do so in different ways and to various degrees with respect to the less well understood implications that LBS may have on outcomes.

Within the CISNET models, the approach taken to modeling LBS was influenced by the mechanics of the models in several ways. Some groups modeled individual life histories in which the simulated individuals are placed both in a situation with screening (intervention simulation) and the counterfactual situation without screening (control simulation); tumor characteristics such as growth rate are the same in both situations. Thus, in the simulation there exists a perfect match between control and intervention simulations, which we term "parallel universes." These models overlay a screening program on the natural history of disease and automatically address the first effect of length biased sampling, namely, screening tends to detect slower-growing tumors with longer preclinical dwell times. However, this tendency alone does not necessarily imply that slower-growing tumors would have better prognosis in both the screening and control simulations, as has been hypothesized.

Erasmus provides a good example of directly modeling LBS by using the "parallel universe" approach mentioned above. In the Erasmus model, if a tumor was detected before it reaches a Weibull-distributed fatal diameter, it would be cured and if detected at a larger size a cancer survival time would be assigned to that individual. In this model, the tumor growth rate was positively correlated with the diameter that a tumor would be clinically detected so that slower growing tumors would tend to be clinically detected at a smaller size than faster-growing tumors (i.e., more likely to be detected before metastasis). Both the Stanford and Rochester models have this property as well. Erasmus takes the further step of accommodating the hypothesis that screening detects disease with better prognosis apart from the benefit of stage shift. In the Erasmus model, cancer survival (if applicable) was negatively correlated to tumor growth rate. Screening preferentially detects slower-growing tumors, and this set of tumors also had a better prognosis in both the screening and the no-screening scenarios since they had a better chance of being detected before metastasis and longer survival if they did metastasize.

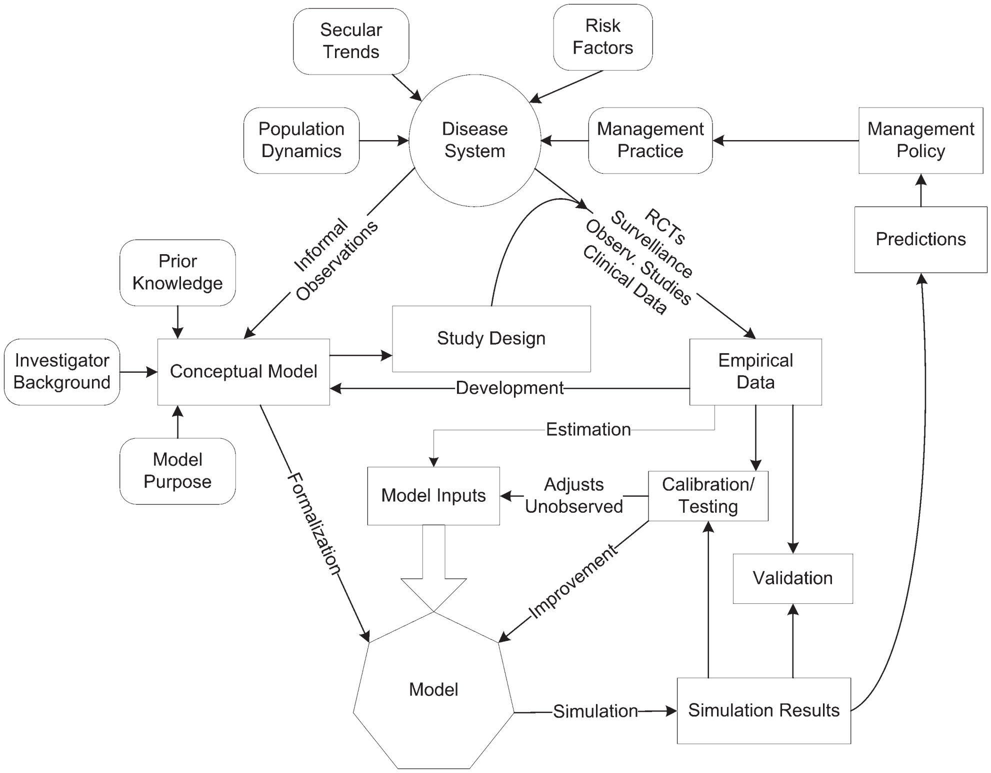

Fig. 1. Relationships and interactions inherent in the modeling process.

Wisconsin captured LBS indirectly. They include a portion of tumors that had a limited malignant potential as a means to address the necessary reservoir of smaller breast cancers that remained undetected until the 1980s with population mammography screening. Limited malignant potential tumors were detected both clinically and through screening, although they were usually detected through screening since they achieved a maximum diameter of 1 cm. By assuming that these tumors had limited malignant potential, the model assured that they would not contribute to mortality in either the screened or unscreened case.

Stanford, Rochester, and Georgetown also use the parallel-universe approach. In these models, screening also will preferentially detect slower-growing tumors. Also, Stanford and Georgetown guarantee that individuals cannot die during their lead time, which gives a small survival benefit apart from stage shift alone. For Stanford and Rochester, screen-detected tumors have a better prognosis than clinically detected tumors overall and within stage since both models assign survival as a function of size at detection, which is linked to growth rate. Screen-detected tumors have a better stage distribution and smaller size within stage. The Georgetown model assigns survival by stage and not size, giving a better overall survival to screen-detected cases, which will have a better stage distribution, but no survival benefit within stage. However, these three models do not link growth rates to survival directly so that there is no link giving slower-growing tumors better survival outside the benefit of stage shift.

The Dana-Farber model is population based in that it mathematically models a population with and without screening. The Dana-Farber model explicitly incorporates the sampling phenomenon in its equations by using the actual distribution of the lead time. The distribution of the lead time is the key variable in the length-biased phenomenon. The DF model assumes that any gain from early detection is a result of a favorable change in the stage distribution because of diagnosing the disease earlier before it transits to more advances stages. As such, they take the position that any increased survival observed in screen-detected cases is the result of a stage shift. The DF model adjusts for the lead time in the eventual survival of screen-detected cases where the survival time is relative to the point in time when the disease would have been diagnosed clinically. This adjustment also depends on the distribution of the lead time.

The University of Texas M. D. Anderson Cancer Center model, also population based, does not include a natural history model and as such does not differentiate between length bias and the benefits of early detection. They include the possibility of length bias by incorporating "beyond stage shift" parameters that govern the benefit of screening beyond what would be expected from stage shift alone *(5)*.

The details of each model's approach to survival from screen-detected disease were described previously *(3–9)*. In most models, the subsequent stage shift (if present) primarily determines an improvement of survival (D, G, R, S, M). Other groups (E, W) model treatment effectiveness as a cure/no-cure process depending on certain characteristics of the cancer at diagnosis (size, age, estrogen receptor [ER] status) and (possibly) treatment standards at the time of diagnosis. Stage shift–driven models attribute a generally more favorable survival distribution to screen-detected cases when screen detection causes the cancer to be detected in an earlier stage. Generally, the survival distribution applied for a particular screen-detected case is estimated from survival as observed in a population-based cancer registry prior to dissemination of screening. The survival used typically depends on some combination of age, stage, ER status, and size of tumor. This conditioning, combined with the fact that screen detection precedes clinical detection, generally yields better survival with screen detection than with clinical detection for the same woman. Survival is further adjusted by treatment effects if treatment was deemed to be in use at the simulated time of diagnosis. Some models apply the screened survival distribution from the time of screen detection. 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). Group R used SEER data to model the effect of treatment as the change in survival over time controlling for age, tumor size, and clinical stage. Thus, treatment for Group R includes not only adjuvant therapy but also better surgical and radiation procedures and improved patient care. Group M also allowed for improvements in breast cancer survival because of other improvements in detection and treatment beyond the interventions considered in the base case question.

Discussion

Challenges Inherent in Model Comparison

Ambiguities in language. Simulation modeling at this interdisciplinary, collaborative level in the health sciences is a relatively nascent endeavor—one that lacks a well-defined vocabulary and a set of common design patterns. These shortcomings, combined with the interdisciplinary makeup of this collaboration with more than a dozen different fields represented, made it necessary to spend a great deal of time interacting to develop a common understanding of key concepts. At the center of this effort was the need to reasonably standardize the vocabulary used in the dialog between collaborators. At times standardization was impossible because of fundamental differences in approach to modeling. When this occurred, the discussion served to “red flag” such terms, and every effort was then made to clarify and give context to the term in question. Some examples of ambiguities that needed to be resolved follow.

Model. The term “model” may refer to something as simple as a linear regression on two variables or to something as complex as a 50 000-line software program with hundreds of inputs running on a cluster of computers. Moreover, models may be composed with other models, yielding hybrids made up of both closed-form analytic models and software-based models. The CISNET collaboration involved a wide range of modeling approaches that required careful qualification to be made when the word “model” was used.

Parameter. To a statistician, a parameter is estimated from sample data. This estimate is subsequently used as a descriptor of the data (e.g., distribution parameters). These estimations are done via statistical techniques that involve an assumed model for the underlying data. That is, a descriptive parameter estimate is considered the result or *output* of a statistical analysis. To a computer scientist, a parameter is an argument passed to a software routine. That is, a parameter in an *input* that determines how the routine will execute. This ambiguity, along with the fact that models that use the technique of Bayesian updating to convert priors (inputs) into posteriors (outputs), adds more blurring of the concept of input and output required the group to provide careful context whenever the word “parameter” was used.

Sojourn time/sensitivity. The term “sojourn time” is typically defined in health science as the duration of the preclinical phase (detectable yet asymptomatic) of disease. However, for simulation modeling an increased level of rigor is required. Since the threshold of detection may vary between different screening modalities, the sojourn time depends on screening modality and in particular on the sensitivity of the screen test used. Sensitivity, in turn, can be understood in terms of estimates from gathered clinical observations of interval cancers, or it may be defined analytically as the “operational sensitivity” the model may use to determine whether a particular test will be simulated as positive or negative in the presence of disease. Finally, the term “sojourn

Table 5. Ancillary data sources

Model*	Additional data sources†	Usage
D	HIP	Estimation of exponential sojourn times
	BCSC	Stage distributions with screening
	Screening Clinical Trials	Mammogram sensitivity and mean sojourn times of preclinical stage and their relation to age
E	Two County Study	Growth rate, survival duration, screening threshold size
	HIP	Screening threshold size
G	HIP, Malmo	Stage-specific dwell times
	Various studies	Sojourn time
	CNBSS et al.	Test sensitivity
M	HIP, CNBSS	Screening effects beyond stage shift
S	None	
R	CNBSS	Natural history and screen parameters (subsequently calibrated via SEER data for U.S. population)
W	WCRS	Initial model building, and incidence calibration
	SEER	Estimation of distant stage survival in prescreening era

*Model group abbreviations: D (Dana-Farber Cancer Center), E (Erasmus University Rotterdam), G (Georgetown University Medical Center), M (University of Texas M. D. Anderson Cancer Center), S (Stanford University), R (University of Rochester), W (University of Wisconsin–Madison).

†Key to additional data source abbreviations. HIP = Health Insurance Plan Project *(16)*; BCSC = Breast Cancer Surveillance Consortium *(17)*; Two County Study = Swedish Two-County Trial *(18)*; Malmo = Malmo mammographic screening trial *(19)*; CNBSS = Canadian National Breast Screening Study *(20)*; WCRS = Wisconsin Cancer Reporting System *(21)*; SEER = Surveillance Epidemiology and End Results *(22)*.

Table 6. Factors affecting benefit from screening (independent of treatment)

		Within stage	
Model*	Primary mechanism	Age	Size
E	Size at diagnosis		✓
S	Stage Shift	✓	✓
M	Stage Shift†	✓	
D	Stage Shift‡	✓	
G	Stage Shift‡	✓	
R	Stage Shift	✓	✓
W	Stage Shift§	✓	

*Model group abbreviations: D (Dana-Farber Cancer Center), E (Erasmus University Rotterdam), G (Georgetown University Medical Center), M (University of Texas M. D. Anderson Cancer Center), S (Stanford University), R (University of Rochester), W (University of Wisconsin–Madison).

†Mortality trends in late stage (III, IV) do not depend on stage shift. This model incorporates a benefit due to screen detection regardless of stage.

‡These models allow substantive benefit only if a stage shift is present with screening.

§This model bases survival benefit on treatment only (see Table 7). However, treatment is more effective for earlier stages and somewhat less effective overall in the elderly to account for historical age bias.

Table 7. Factors affecting benefit from treatment (independent of screening status)

Model*	ER status	Age	Calendar year
E		✓	
S	✓	✓	
M	✓	✓	✓†
D	✓	✓	✓
G	✓	✓	
R		✓	✓†
W‡	✓	✓	✓

*Model group abbreviations: D (Dana-Farber Cancer Center), E (Erasmus University Rotterdam), G (Georgetown University Medical Center), M (University of Texas M. D. Anderson Cancer Center), S (Stanford University), R (University of Rochester), W (University of Wisconsin–Madison).

†Captures not only adjuvant therapy but also better procedures and generally improved patient care.

‡This group models treatment as a cure/no-cure process; probability of cure depends on estrogen receptor (ER) status, age, and treatment type (which depends on calendar year).

time” has a more general meaning in the simulation literature in which it is taken to mean the time spent in some state tracked by the simulation. Thus the term “sojourn time” requires further modification to convey meaning in the context of simulation modeling in health sciences.

Challenges due to the nature of modeling. Modeling is an iterative process with many upstream, dynamic dependencies. These models are thus never complete, as they embody a microcosm of science itself with continual revision being made in the face of new knowledge or applications. At the outset, observations are made, data are collected, and preliminary analyses are carried out—often guided by previous work and prevailing hypotheses generated by clinical observation; theories are developed and models are built to test and compare those theories in hypothetical situations. Any change in the precursors to the model (observation, data, analysis technique, or theories) can necessitate a change in the approach to modeling the system. The models themselves can form the catalyst for change in that model results may inform future policy direction and collections of empirical data. This brief description of the information flow vastly oversimplifies the case, as in reality the situation is far from linear and there are many internal feedback loops. The actual interactions are better, albeit still incompletely, represented in Fig. 1.

The backdrop of the highly dynamic, related process of building models raises challenges both in the comparison of models and model results and the documentation of model structure. The CISNET Breast Cancer program sought to mitigate these difficulties by standardizing the analysis target, key inputs, and the documentation approach across all models. This level of standardization was an important aspect of the success of the project. Standards were put into place for presentation of model output and documentation. These standards allowed those tasked with evaluating results to rapidly determine the likely reasons for differences between the models and determine whether they were based in valid differences in approach and/or assumptions or if in fact the definitions of the standards needed to be clarified.

Challenges due to uncertainty in structure and inputs. As described earlier, the models involved in this collaboration often took different approaches and philosophies to model building. Although it is unlikely there are obvious errors, given the unobservable nature of the cancer disease process, it is doubtful that any particular model is correct. There will always remain a certain degree of uncertainty in modeling structure. Similarly, although efforts were made to standardize as much as possible the raw materials each model used in this analysis, different philosophies and approaches necessitated that different datasets be used at times (see Table 5). These different datasets imply different uncertainties in the parameters estimated from them. Although it is possible to quantify the uncertainty inherent in model results due to parameter uncertainty, it is much more difficult to quantify the uncertainty due to model structure. The standardized collaboration used by the CISNET group to some extent enables us to look at structure uncertainty in a qualitative way (see Tables 6 and 7); the interplay between parameter uncertainty and structure uncertainty remains largely unknown. This difficulty is not unique to the CISNET collaboration. Indeed, this difficulty exists in any comparison of results in the modeling literature. The collaborative nature of this effort offered an unprecedented opportunity for discourse on differing approaches, which lent important contextual clues when results were compared. Much of this dialog would have been impossible if each group worked independently to answer similar or related questions. This collaboration developed both models and tools and processes by which models can be compared.

References

(1) Feuer EJ. Modeling the impact of adjuvant therapy and screening mammography on U.S. breast cancer mortality between 1975 and 2000: introduction to the problem. J Natl Cancer Inst Monogr 2006;36:2–6.

(2) National Cancer Institute. CISNET Model Profiles. Available at: http://cisnet.cancer.gov/profiles.

(3) Tan SYGL, van Oortmarssen GJ, de Koning JH, Boer R, Habbema JDF. The MISCAN-Fadia continuous tumor growth model for breast cancer. J Natl Cancer Inst Monogr 2006;36:56–65.

(4) Plevritis SK, Sigal BM, Salzman P, Rosenberg J, Glynn P. A stochastic simulation model of U.S. breast cancer mortality trends from 1975 to 2000. J Natl Cancer Inst Monogr 2006;36:86–95.

(5) Berry DA, Inoue L, Shen Y, Venier J, Cohen D, Bondy M, et al. Modeling the impact of treatment and screening on U.S. breast cancer mortality: a Bayesian approach. J Natl Cancer Inst Monogr 2006;36:30–6.

(6) Lee S, Zelen M. A stochastic model for predicting the mortality of breast cancer. J Natl Cancer Inst Monogr 2006;36:79–86.

(7) Mandelblatt J, Schechter CB, Lawrence W, Yi B, Cullen J. The SPECTRUM population model of the impact of screening and treatment on U.S. breast cancer trends from 1975 to 2000: principles and practice of the model methods. J Natl Cancer Inst Monogr 2006;36:47–55.

(8) Hanin LG, Miller A, Zorin AV, Yakovlev AY. The University of Rochester model of breast cancer detection and survival. J Natl Cancer Inst Monogr 2006;36:66–78.

(9) Fryback DG, Stout NK, Rosenberg MA, Trentham-Dietz A, Kuruchittham V, Remington PL. The Wisconsin Breast Cancer Epidemiology simulation model. J Natl Cancer Inst Monogr 2006;36:37–47.

(10) Cronin KA, Feuer EJ, Clarke LD, Plevritis SK. Impact of adjuvant therapy and mammography on U.S. mortality from 1975 to 2000: comparison of mortality results from the CISNET breast cancer base case analysis. J Natl Cancer Inst Monogr 2006;36:112–21.

(11) Habbema JDF, Tan SYGL, Cronin KA. Impact of mammography on U.S. breast cancer mortality, 1975–2000: are intermediate outcome measures informative? J Natl Cancer Inst Monogr 2006;36:105–11.

(12) Mariotto AB, Feuer EJ, Harlan LC, Abrams J. Dissemination of adjuvant multiagent chemotherapy and tamoxifen for breast cancer in the United States using estrogen receptor information: 1975–1999. J Natl Cancer Inst Monogr 2006;36:7–15.

(13) Rosenberg MA. Competing risks to breast cancer mortality. J Natl Cancer Inst Monogr 2006;36:15–9.

(14) Holford TR, Cronin KA, Mariotto AB, Feuer EJ. Changing patterns in breast cancer incidence trends. J Natl Cancer Inst Monogr 2006;36:19–25.

(15) Cronin KA, Mariotto AB, Clarke LD, Feuer EJ. Additional common inputs for analyzing impact of adjuvant therapy and mammography on U.S. mortality. J Natl Cancer Inst Monogr 2006;36:26–9.

(16) Shapiro S. Periodic screening for breast cancer: the HIP Randomized Controlled Trial. Health Insurance Plan. J Natl Cancer Inst Monogr 1997;22: 27–30.

(17) Ballard-Barbash R, Taplin SH, Yankaskas BC. Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. AJR Am J Roentgenol 2004;169:1001–8.

(18) Tabar L, Vitak B, Chen HH, Duffy SW, Yen MF, Chiang CF, et al. The Swedish Two-County Trial twenty years later. Updated mortality results and new insights from long-term follow-up. Radiol Clin North Am 2000;38: 625–51.

(19) Andersson I, Aspegren K, Janzon L, Landberg T, Lindholm K, Linell F, et al. Mammographic Screening and mortality from breast cancer: the Malmo mammographic screening trial. BMJ 1988;297:943–8.

(20) Miller AB, To T, Baines CJ, Wall C. The Canadian National Breast Screening Study-2: 13-year results of a randomized trial in women age 50–59 years. J Natl Cancer Inst 2000;92:1490–9.

(21) Wisconsin Cancer Incidence and Mortality, 1999. Madison (WI) Bureau of Health Information, Division of Health Care Financing, Wisconsin Department of Health and Family Services; 2002.

(22) Ries LA, Eisner MP, Kosary CL, Hankey BF, Miller BA, Clegg L, et al. (eds). SEER Cancer Statistics Review, 1975–2002, National Cancer Institute. Bethesda, MD. Available at: http://seer.cancer.gov/csr/1975_2002/, based on November 2004 SEER data submission, posted to the SEER Web site 2005.

(23) National Cancer Institute. CISNET Input Parameter Generator Interfaces. Available at: http://cisnet.cancer.gov/interfaces.

(24) Joensuu H, Lehtimaki T, Holli K, Elomaa L, Turpeenniemi-Huianen T, Kataja V, et al. Risk of distant recurrence of breast cancer detected by mammography screening or other methods. JAMA 2004;292:1064–73.