SIMULATION MODELING TO ENHANCE COMPARATIVE EFFECTIVENESS RESEARCH IN CANCER

Jeanette Birnbaum

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Ruth Etzioni, Chair

Paul Hebert

Steven Zeliadt

Joshua Carlson

Program Authorized to Offer Degree:

Health Services

University of Washington

ABSTRACT

Simulation modeling to enhance comparative effectiveness research in cancer

Jeanette Birnbaum

Chair of the Supervisory Committee:
Ruth Etzioni, Affiliate Professor
Department of Health Services

Policy-makers need compelling evidence of the relative benefits and harms or "comparative effectiveness" (CE) of new interventions in order to guide clinical practice and prioritize future research. In cancer control, the primary indicator of intervention efficacy is reduction in cancer deaths. However, CE studies often do not follow subjects to the time of death due to time and resource constraints. Even when CE studies do reach the time of death, the results may quickly become obsolete due to advances in technology that occur during the long study period. Models can integrate data from various studies to project a coherent picture of disease progression from intervention to death. The goal of this research is to construct a novel modeling framework to translate existing CE results into their implications for deaths prevented in three cancer control settings: 1) using diagnostic tests to target treatment 2) using new biomarkers to enhance screening 3) screening under evolving treatment technology. In each setting, we build a transparent model that can integrate data from multiple studies to project deaths prevented. We then apply the model to answer an important current question in that

setting. In the first setting, we evaluate whether a new diagnostic test for breast cancer used to identify patients most likely to benefit from treatment actually changes the risk of cancer death by changing the rate at which patients get treated. In the second setting, we consider whether a new biomarker for prostate cancer that changes the sensitivity and specificity of prostate cancer screening actually changes the risk of prostate cancer death and overdiagnosis. In the third setting, we address whether advances in breast cancer treatments over the last two decades have implications for interpreting screening mammography trials conducted before these treatments were available. In each case, we use published data from a CE study as a starting point and build on this to extrapolate the observed CE results into a projection of deaths prevented.

ACKNOWLEDGEMENTS

I would like to express a heartfelt thanks to my dissertation committee, whose willingness to invest in me made all the difference. Ruth Etzioni took me under her wing as a research assistant and patiently trained me to be a modeler. She helped me find a path within health services research that fit both my love of quantitative research and my interest in health policy. Under her mentorship, I not only gained a whole new analytic skill set but also learned to be a sharper thinker and a better writer. Her faithfulness to her principle "family and friends first" was also a wonderful example to me. Paul Hebert and Steve Zeliadt helped me find a clear path through the maze of research and training options in which students can easily get lost. Their unflagging faith in me kept me motivated, and their generally cheerful outlook reminded me that research is fun! Josh Carlson was always available when I needed guidance in questions of cancer genomics, cost and outcomes methodology, and navigating peer review. My GSRs Adrian Dobra and Martina Morris encouraged me in my dissertation and were a pleasure to work with in other settings as well.

I am also thankful to the members of the Etzioni lab and others at the Fred Hutchinson Cancer Research Center and the University of Washington who supported me over the last four years. Roman Gulati's door was always open to me despite his busy schedule, and I attribute most of my technical skills development to him. Jing Xia and Lurdes Inoue were also always ready to answer my statistical questions. Leslie Mallinger made the final year of the CANTRANce project much more productive, high-quality and enjoyable through her work and friendship. Mark Mason's work on the CANTRANce interface was invaluable, as was his willingness to continue to contribute time after the project officially ended. Dave Grembowski and Bonnie Duran enthusiastically mentored me in other research interests along the way.

I would not have gotten very far at all in this endeavor without my friends and family. My fellow student moms showed me how to balance family and research and were wonderfully supportive. My husband Kevin bore every up and down with incredible patience and love. My son Leo continually reminded me that life is amazing and that the present is precious. New baby, you made the last 6 months a lot harder, but you also gave me the most motivating deadline possible!

I would also like to acknowledge the financial support I received through my research assistantship with Ruth Etzioni, my teaching assistantship with the Department of Health Services, and the Comparative Effectiveness Dissertation Award from the Pharmaceutical Outcomes Research and Policy Program.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Background

Demand for comparative effectiveness research (CER) has grown swiftly over the last 5 years, following its spotlight in a 2008 Institute of Medicine report and funding provided by the 2009 American Recovery and Reinvestment Act that established the Patient-Centered Outcomes Research Institute.[1–3] These milestones signal widespread recognition that evidence on the relative harms and benefits of interventions is needed for modern medical decision-making, given the dynamic landscape of available health care technologies and rising health care costs.[4]

Recent dialogue on CER has consequently focused on developing methods to better generate and synthesize the breadth of medical data being collected through electronic medical records, patient registries and other sources. CER in the twenty-first century must be rapid, relevant, and efficient [5]. Randomized controlled trials (RCTs), long considered the gold-standard design for studying interventions, have been called "ill-suited" for CER due to their often-atypical patient populations and care settings and long timeframe.[6,7] Design and analytic changes to the classic RCT are thus emerging CER topics, along with methods for strengthening observational studies.[6,8–13]

Certain limitations of randomized trials and observational studies are difficult to overcome, however, and alternative CER methods are necessary to supplement these traditional study designs.[14] Even with adaptive designs or better use of electronic medical records and administrative databases, RCTs and observational studies are victims of timing: they must strike a balance between studying intermediate endpoints and conducting lengthy follow-up.[11,15]

Both options have limitations that put rapid research and efficiency at odds with relevance. Studies that produce results rapidly necessarily exclude long-term outcomes such as

mortality, whereas studies with long follow-up cannot account for technological advances occurring during the study period and thus sacrifice relevance. The efficiency of traditional study designs also decreases with length of follow-up. Examples abound of studies on both ends of the timing spectrum. More recent technologies such as robotic prostatectomy and high-intensity ultrasound therapy appear promising, but based on short follow-up.[16,17] Breast and prostate cancer screening trials have decades of follow-up, but the observed mortality reflects treatment efficacies and patterns of treatment that are now outdated.[18]

Modeling attempts to address these limitations of CE studies by synthesizing available CE results into a more complete picture of intervention effectiveness. Models can estimate long-term outcomes while taking into account the most contemporary data available. The tension between rapid, relevant, and efficient CER disappears in modeling research because models can answer current questions in CER quickly, using only programming and computing resources. The US Preventive Services Task Force recently used modeling studies investigating hundreds of potential screening protocols in breast and colorectal cancers to guide their screening recommendations.[19,20] Modeling is increasingly being recognized as an underutilized tool for CER.[21]

Modeling studies are timely and efficient compared to traditional studies, but barriers to the use of models for CER remain. One is their perception as a “black box” with hidden, built-in assumptions.[22] While some models are necessarily complex, there are CER questions that may be addressed by simpler models with more accessible components. Another barrier is that model development requires time and resources, and existing models in the literature are typically highly application-specific. Failure to recognize common elements across models so that they can be transferred to new applications has led to vast amounts of time and resources being devoted to customized model-building.

Simple models that can be transferred across applications and have accessible components can both increase the acceptability and utilization of modeling for CER and, in turn, provide timely evidence for policy-making. This dissertation presents three such models that answer outstanding CER questions in cancer screening and diagnosis.

Specific Aims

Aim 1: Project the impact of testing for biomarkers to target cancer treatment on mortality, quality-adjusted life-years, and costs

A topic of great interest within the new area of personalized medicine is the use of biomarkers at the time of diagnosis to target treatment towards patients who are most likely to benefit from them. CE studies evaluating these "diagnostic" biomarkers typically estimate the effect of testing on treatment distributions rather than on mortality. In this aim we construct a model to project the mortality and quality-adjusted mortality implications of testing for a diagnostic biomarker compared to no testing. We use this model to do a novel projection of how changes in treatment recommendations due to the "21-gene recurrence score," a new molecular assay for breast cancer patients, would translate to impacts on mortality and costs. We additionally provide an accompanying user interface to facilitate access to the model.

Aim 2: Project the impact of novel cancer screening tests on mortality and overdiagnosis

Historically, less than 1% of cancer screening tests initially found to have good diagnostic properties have reached clinical practice. This is because promising estimates of sensitivity and specificity at early stages of diagnostic test development do not always translate into significant mortality benefits at a reasonable cost. Early projections from diagnostic properties to the mortality impact of novel screening tests are needed to prioritize the development of initially promising tests. In this aim we build a model to project the mortality and overdiagnosis implications of introducing a novel screening test compared to standard or no screening. We

this model to do the first impact evaluation of introducing Prostate Cancer Antigen 3 (PCA3) as a new screening test for prostate cancer in addition to Prostate-Specific Antigen (PSA).

Aim 3: Project the impact of treatment advances on the mortality results of cancer screening trials

In order to observe the impact of screening on mortality, cancer screening trials must span decades. Advances in the efficacy and diffusion of treatment for early or advanced-stage cancers during the long follow-up of screening may affect the mortality impact of more contemporary screening programs. In this aim we construct a model that updates the impact of a screening test as observed in a published trial to account for changes in treatment use and efficacy. We apply this model to explore whether treatment advances in breast cancer that postdate mammography screening trials have changed the mortality impact of screening.

Impact

The models in this dissertation each serve a dual purpose: one, to provide a novel, re-usable framework for projecting the mortality impact of a cancer intervention; and two, to provide timely new insight into a topical CE issue in cancer. Each model achieves the first, general purpose through a particular mechanism. In Aim 1, the model pushes the forefront of reproducible research through a companion Windows application interface and tutorial that allows naïve CE researchers to run the model in a guided setting. While modeling is not uncommon surrounding questions of new diagnostic tests, such direct access is quite new. Future applications to other diagnostic tests in breast or other cancers would be extremely efficient and limited only by the availability of the required input CE data. The Aim 2 model is incrementally constructed to provide a template for the evaluation of any new screening biomarker with CE data on its diagnostic properties. This paves the way for the as-yet unrecognized, valuable role of modeling

in prioritizing among the many initially promising biomarkers that have been identified. The model itself is specific to prostate cancer but could flexibly study additional screening biomarkers. In Aim 3, the approach and the model itself are both transferrable not only to other screening tests in other cancers, but also to additional variations of the same CE dilemma. For example, the model could investigate the impact of treatment advances that occur differentially across the population on disparities.

The specific applications of each model complete each Aim by contributing novel insights to CER in cancer screening and diagnosis. Aim 1 provides a novel projection of the comparative effectiveness of testing with the 21-gene recurrence score by extending a CE study of actual physician practice patterns in the presence and absence of the test. Previous models have been based on assumptions of how the test is used in clinical practice rather than empirical data.[23–27] Aim 2 provides the first projections of how PCA3 will impact prostate cancer outcomes if it is introduced as a new screening tool. A recent AHRQ panel review of PCA3 identified such projections as one of the immediate priorities in PCA3 research.[28] Aim 3 highlights that the benefit of screening is contingent upon treatment utilization and efficacy, a fact often neglected or misunderstood when screening study results are interpreted. The mammography study provides the first quantitative framework to inform recent debates regarding the impact of mammography screening in the context of contemporary therapies.[29] As a whole, the dissertation enhances CER in cancer by providing new tools and directions for future modeling research along with insights into current CE questions.

References

1. Sox, H. C. & Greenfield, S. Comparative Effectiveness Research: A Report From the Institute of Medicine. *Ann. Intern. Med.* **151,** 203–205 (2009).
2. Selby, J. V., Beal, A. C. & Frank, L. The Patient-Centered Outcomes Research Institute (PCORI) national priorities for research and initial research agenda. *JAMA J. Am. Med. Assoc.* **307,** 1583–1584 (2012).
3. Golub, R. M. & Fontanarosa, P. B. Comparative effectiveness research: relative successes. *JAMA J. Am. Med. Assoc.* **307,** 1643–1645 (2012).
4. Lyman, G. H. & Levine, M. Comparative effectiveness research in oncology: an overview. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **30,** 4181–4184 (2012).
5. Garber, A. M. & Tunis, S. R. Does Comparative-Effectiveness Research Threaten Personalized Medicine? *N. Engl. J. Med.* **360,** 1925–1927 (2009).
6. Luce, B. R. *et al.* Rethinking Randomized Clinical Trials for Comparative Effectiveness Research: The Need for Transformational Change. *Ann. Intern. Med.* **151,** 206–209 (2009).
7. Sullivan, P. & Goldmann, D. The promise of comparative effectiveness research. *JAMA J. Am. Med. Assoc.* **305,** 400–401 (2011).
8. Dreyer, N. A. *et al.* Why observational studies should be among the tools used in comparative effectiveness research. *Health Aff. Proj. Hope* **29,** 1818–1825 (2010).
9. Johnson, M. L., Crown, W., Martin, B. C., Dormuth, C. R. & Siebert, U. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report--Part III. *Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res.* **12,** 1062–1073 (2009).
10. Cox, E. *et al.* Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: the International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report--Part II. *Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res.* **12,** 1053–1061 (2009).
11. Carpenter, W. R., Meyer, A.-M., Abernethy, A. P., Stürmer, T. & Kosorok, M. R. A framework for understanding cancer comparative effectiveness research data needs. *J. Clin. Epidemiol.* **65,** 1150–1158 (2012).
12. Ginsburg, G. S. & Kuderer, N. M. Comparative effectiveness research, genomics-enabled personalized medicine, and rapid learning health care: a common bond. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **30,** 4233–4242 (2012).
13. Hershman, D. L. & Wright, J. D. Comparative effectiveness research in oncology methodology: observational data. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **30,** 4215–4222 (2012).
14. Armstrong, K. Methods in comparative effectiveness research. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **30,** 4208–4214 (2012).
15. Garnett, G. P., Cousens, S., Hallett, T. B., Steketee, R. & Walker, N. Mathematical models in the evaluation of health programmes. *Lancet* **378,** 515–525 (2011).
16. Robertson, C. *et al.* Relative effectiveness of robot-assisted and standard laparoscopic prostatectomy as alternatives to open radical prostatectomy for treatment of localised

prostate cancer: a systematic review and mixed treatment comparison meta-analysis. *BJU Int.* (2013). doi:10.1111/bju.12247
17. Chan, A. C. Y. *et al.* Survival analysis of high-intensity focused ultrasound therapy versus radiofrequency ablation in the treatment of recurrent hepatocellular carcinoma. *Ann. Surg.* **257,** 686–692 (2013).
18. Gøtzsche, P. C. & Jørgensen, K. J. Screening for breast cancer with mammography. *Cochrane Database Syst. Rev.* **6,** CD001877 (2013).
19. Zauber, A. G. *et al.* Evaluating test strategies for colorectal cancer screening: a decision analysis for the U.S. Preventive Services Task Force. *Ann. Intern. Med.* **149,** 659–669 (2008).
20. Mandelblatt, J. S. *et al.* Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. *Ann. Intern. Med.* **151,** 738–747 (2009).
21. Glasgow, R. E. *et al.* Comparative effectiveness research in cancer: what has been funded and what knowledge gaps remain? *J. Natl. Cancer Inst.* **105,** 766–773 (2013).
22. Mandelblatt, J. *et al.* Building better models: if we build them, will policy makers use them? Toward integrating modeling into health care decisions. *Med. Decis. Mak. Int. J. Soc. Med. Decis. Mak.* **32,** 656–659 (2012).
23. Hornberger, J., Chien, R., Krebs, K. & Hochheiser, L. US Insurance Program's Experience With a Multigene Assay for Early-Stage Breast Cancer. *J. Oncol. Pract. Am. Soc. Clin. Oncol.* **7,** e38s–45s (2011).
24. Hornberger, J., Cosler, L. E. & Lyman, G. H. Economic analysis of targeting chemotherapy using a 21-gene RT-PCR assay in lymph-node-negative, estrogen-receptor-positive, early-stage breast cancer. *Am. J. Manag. Care* **11,** 313–324 (2005).
25. Lyman, G. H., Cosler, L. E., Kuderer, N. M. & Hornberger, J. Impact of a 21-gene RT-PCR assay on treatment decisions in early-stage breast cancer: an economic analysis based on prognostic and predictive validation studies. *Cancer* **109,** 1011–1018 (2007).
26. Reed, S. D., Dinan, M. A., Schulman, K. A. & Lyman, G. H. Cost-effectiveness of the 21-gene recurrence score assay in the context of multifactorial decision making to guide chemotherapy for early-stage breast cancer. *Genet. Med. Off. J. Am. Coll. Med. Genet.* (2012). doi:10.1038/gim.2012.119
27. Vanderlaan, B. F., Broder, M. S., Chang, E. Y., Oratz, R. & Bentley, T. G. K. Cost-effectiveness of 21-gene assay in node-positive, early-stage breast cancer. *Am. J. Manag. Care* **17,** 455–464 (2011).
28. Gutman, S. I., Oliansky, D. M., Belinson, S. & Aronson, N. *PCA3 Testing in the Diagnosis and Management of Prostate Cancer: Future Research Needs: Identification of Future Research Needs From Comparative Effectiveness Review No. 98*. (Agency for Healthcare Research and Quality (US), 2013). at <http://www.ncbi.nlm.nih.gov/books/NBK143597/>
29. Printz, C. Mammogram debate flares up: Latest breast cancer screening study fuels controversy. *Cancer* **120,** 1755–1756 (2014).

CHAPTER 2: COMPARATIVE EFFECTIVENESS OF BIOMARKERS TO TARGET CANCER TREATMENT: MODELING IMPLICATIONS FOR SURVIVAL AND COSTS

Introduction

Over the last decade, advances in genomic research have produced biomarkers to target cancer treatments to a patient's unique molecular characteristics.[1–4] Testing patients for biomarkers that either predict response to particular cancer therapies or supply prognostic information can support tailored treatment decisions that improve outcomes and/or reduce toxicity. Established examples of such biomarkers include the Her-2/Neu oncogene that predicts positive response to trastuzumab for breast cancer and the KRAS mutations that indicate resistance to anti-EGFR treatments for metastatic colorectal cancer.[5] In both of these examples, randomized controlled trials (RCTs) of patients who are positive for the biomarkers in question have demonstrated better survival with tailored treatment.[4,6–10]

In contrast, emerging biomarkers may diffuse into clinical practice long before survival data become available. Depending on the cancer, more than a decade may pass between early-stage studies of a diagnostic biomarker and an RCT that confirms benefit of testing and tailored treatment on survival, if such a study is even done.[11] In the interim, comparative effectiveness (CE) studies typically collect data on how testing for a biomarker impacts an intermediate endpoint such as treatment recommendations.[12–17] While data on treatment recommendations can suggest that testing has affected treatment, these data do not address the ultimate question of whether testing reduces cancer mortality.

Quantitative estimates of the mortality impact of testing are needed to assess the value of these biomarkers, which we will call "diagnostic biomarkers" given their use at the time of diagnosis to inform treatment choices. One way to meet this need is to conduct a complementary modeling study that translates changes in treatment recommendation into projected effects on mortality. This strategy of using models for comparative effectiveness

research (CER) when observational data are unavailable has been highlighted as both important and underutilized in cancer research.[18,19] Moreover, the fast pace of biomarker development has created a demand for CER that cannot be fully met by observational studies: when competing diagnostic biomarkers emerge, there may be more head-to-head comparisons than can be feasibly investigated. This environment is a perfect opportunity for modeling to supply otherwise unattainable comparative effectiveness results.

A practical obstacle to modeling the impact of diagnostic biomarkers is that the investigators who conduct CE studies of how diagnostic biomarkers impact treatment typically do not have the additional resources to conduct the complementary modeling study translating their results to mortality. On the other hand, research groups with modeling expertise often conduct modeling studies using aggregate data from the literature because they are not connected with investigators who have primary data. In addition, many modeling studies are highly application-specific and customized to the particular question at hand, preventing them from being quickly re-applied as new diagnostic biomarkers emerge.

To address this gap, we have developed a system of statistical simulation models for Cancer Translation of Comparative Effectiveness Research (CANTRANce) that facilitates the modeling of the mortality impact of interventions, given CE studies of typical intermediate endpoints, i.e. the common pre-mortality outcomes used to demonstrate the effectiveness of interventions. CANTRANce can model the mortality impact of several intervention-intermediate endpoint pairs: preventive interventions and disease incidence; screening tests and sensitivity and specificity of detection or stage distribution of detected cancers; diagnostic biomarkers and treatment distributions; and treatments and disease recurrence. Users enter data from a CE study of how an intervention impacts an intermediate endpoint, and the modeling system prompts them for additional data from the literature or other studies. The system then extrapolates from the intermediate endpoint to the projected impact of the intervention on mortality. In this paper, we introduce the Diagnostic Biomarker module of CANTRANce and use

it to evaluate a diagnostic biomarker for breast cancer. We show how CANTRANce can flexibly model the impact of a new diagnostic biomarker on mortality, as well as on quality-adjusted life-years and costs.

Methods

CANTRANce Framework

In all CANTRANce models, we use an individual-level microsimulation approach in which individuals transition through health states, and transitions occur in continuous time rather than at specified time intervals.[20,21] The models simulate a virtual population and project outcomes for two scenarios, one in which individuals do receive the intervention and one in which they do not. The impact of the intervention on the intermediate endpoint is based on a CE study ending at that endpoint. We model each person's subsequent series of health states or "life history" by projecting the dates of key events such as disease progression and mortality.[20,21]

The Diagnostic Biomarker (Diagnostics) module of CANTRANce evaluates the impact of testing for a new diagnostic biomarker to target treatment (Figure 2.1). The model compares two scenarios: 1) all individuals receive a one-time testing intervention using the diagnostic biomarker and 2) no individuals receive testing. The foundation for the model is a CE study of how the testing intervention impacts the intermediate endpoint, treatment recommendations. Based on this CE study, the testing intervention and no testing scenarios have different distribution of two possible treatments: a standard treatment, and a tailored treatment whose indication changes after test results are known. In order to project to mortality, the model prompts the user for the necessary additional information on expected cancer mortality following each treatment option. Alternatively, the user may project to mortality through a second intermediate endpoint of disease recurrence.

The precise inputs and steps involved in the projection depend on certain choices available to the user. These elements are described in detail below from the perspective of a user running the model via the user-friendly Windows application interface, downloadable from http://www.fhcrc.org/cantrance. The interface prompts the user for the necessary inputs and then runs the model locally using an open-source R package, *cantrance* (v 1.1). The results and open-source code for the model are stored on the user's local machine.

General modeling selections

The user must specify the number of simulations and years at which to evaluate survival. Each simulation represents one study linking testing to treatment and mortality, and multiple simulations capture the stochastic uncertainty due to the simulation process.[22] Users can also define multiple follow-up time points for the projection, e.g. 10-year as well as lifetime follow-up.

Constructing the virtual population

The Diagnostics model begins by generating a virtual population that mimics the population in the CE study. At minimum, each individual in the virtual population needs to be assigned a gender, age at treatment, and either 1) treatment recommended under the testing intervention versus under no testing, for paired CE studies in which treatment was assessed for the same patients under both scenarios, or 2) treatment recommended and an intervention indicator, for unpaired CE studies in which the testing intervention versus no testing was studied using different patients. Other relevant covariates may also be included. The two options for generating this population are:

1. *Specify covariate summaries*. If the user only has population-level summary statistics for covariates reported in the CE study publication, the Diagnostics model can use those to generate a virtual study population for each simulation in which individuals have covariate values that are consistent with the published data at the population level. Both marginal and

joint distributions of categorical and continuous covariates can be accommodated. The Diagnostics model will assume independence between covariates without specified joint distributions.

2. *Bootstrap from individual-level data*. If individual-level data from the CE study are available to the user, the entire CE study population can be entered into the model. For each simulation, the model will bootstrap from this population to define the virtual study population. Bootstrapping preserves all the interrelations between covariates in the original study. While the default bootstrap is unweighted, the user may alternatively specify weights for the resampling in order to model the impact of testing on a population with a particular covariate distribution.

Modeling treatment received

The virtual population includes an indicator of treatment recommended under each scenario, and this value can be used directly. Alternatively, if individual-level data are available, the user may choose to run a logistic regression of treatment received, separately for each scenario, based on available predictors in the data. For the testing intervention scenario, the test result is a required predictor. In each simulation, CANTRANce fits the regression and uses the results to simulate treatment recommended under each scenario. If the CE study is very small, the regression-based treatment distributions for the virtual population may be more stable than the observed values.

Modeling time from treatment to mortality

This step projects the dates of cancer and other-cause mortality for each virtual individual. The cause of death is determined by the event that occurs first.

The Diagnostics model projects times to event using a single, user-specified survival statistic, e.g., an event rate or k-year survival. This statistic is translated into an exponential

survival curve from which the model simulates times to event for each individual. Hazard ratios may be specified to modify this curve for particular covariate values.

It is important to note that the survival statistic is interpreted as a *net* statistic, that is, in the absence of other-cause death. Users may find some net statistics reported directly in the literature. In particular, relative survival is an appropriate net statistic[23] routinely reported by SEER Cancer Statistics Review.[24] The Surveillance, Epidemiology and End Results (SEER) database[25] provides another venue for deriving net survival statistics. Alternatively, a composite statistic such as overall survival or disease-free survival can be used to approximate a net statistic when the risk of death from other causes is relatively low in the study cohort (e.g., in a younger cohort).

The user may choose to use disease recurrence as an intermediate endpoint within the time to mortality projection. If so, the Diagnostics model requires one survival statistic for the time from treatment to recurrence, and another statistic for time from recurrence to mortality. Otherwise, the user specifies one survival statistic describing time from treatment to mortality. In either case, the user must specify at least one hazard ratio describing the benefit of tailored versus standard treatment on the event (recurrence and/or mortality). Additional hazard ratios may be specified if survival or the benefit of tailored treatment varies by other covariates.

Age at other-cause death is simulated from US cohort life tables.[26] A hazard ratio on the life table may be specified if the CE study population is thought to have all-cause survival that differs from the general population.[27]

Modeling quality of life and cost

Because some diagnostic biomarkers primarily function to spare patients unnecessary treatment and thus improve quality of life, the Diagnostics model also allows users to specify annual costs and utility weights for key periods during the life history: the first year of treatment, subsequent years until mortality, and the year of death. If recurrence is used in modeling

survival, cost and utilities can be further partitioned by the onset of recurrence. All costs and utility weights can vary by tailored versus standard treatment. The user may also specify a discount rate to discount future gains to their present value.[28]

Quantifying uncertainty

Each run of the Diagnostics model performs multiple simulations using the same parameter set.[20] Stochastic (i.e., Monte Carlo) uncertainty is quantified by summarizing results across simulations using the mean as the point estimate and the 2.5% and 97.5% quantiles as the 95% uncertainty interval. Sensitivity to parameter values can be easily investigated by changing parameters through the user interface. Formal one-way and probabilistic sensitivity analyses that systematically investigate parameter sets are also possible outside of the interface, using some customized R programming. For sensitivity analyses, we again describe uncertainty using the 2.5% and 97.5% quantiles across runs to define the 95% uncertainty interval.

Case Studies

The Diagnostics model is designed to be a high-quality, flexible tool to project long-term outcomes following testing with a diagnostic biomarker. We developed two case studies that highlight these features. The first case compares the Diagnostics model's performance to that of a previously developed customized model, and the second case provides novel CE projections in a setting where there is no existing model.

Both cases are founded on CE studies of how a testing intervention using a 21-gene recurrence score (RS) for breast cancer impacts treatment recommendations for adjuvant chemotherapy plus hormone therapy (tailored treatment) versus hormone therapy alone (standard treatment). The RS classifies patients as being at low, intermediate or high risk of disease recurrence. The CE studies describe how physician treatment recommendations change with and without RS knowledge. The Diagnostics model translates these changes in

treatment to their impact on mortality, quality-adjusted life years and costs. In both cases, we discount outcomes using the standard rate of 3%.[29]

Case 1: Replication of a customized model

In Case 1 we replicated projections made by a customized model developed by Reed et al to compare using the RS to guide treatment (the testing intervention, or "with RS") to NCCN-guideline-based treatment (no testing, or "without RS").[30] Reed et al used a Markov cohort decision model to project outcomes for these two scenarios based on a published CE study in which medical oncologists made two treatment recommendations for node-negative, ER-positive patients, first without the patients' RS and then with the RS.[12,31] Because the model used published CE data, no individual-level data were available and we thus used the covariate summaries option of the Diagnostics model to replicate the CE study population reported in Reed et al. We projected time to mortality, total costs from the societal perspective, and quality-adjusted life-years using recurrence as a second intermediate endpoint (Table 2.1). We additionally replicated the probabilistic sensitivity analysis (PSA) using the same standard errors and distributional assumptions (Table 2.1). As a sensitivity analysis, we investigated the impact of customizing the Diagnostics model to match the Reed model on two final elements: their period life table (versus our cohort life table) and their assumption that recurrence is possible only in the first 10 years (versus our unlimited possibility of recurrence).

Case 2: Novel projection of the impact of diagnostic testing

In Case 2, we did a novel projection of the impact of the RS using individual-level data from a CE study by Ademuyiwa et al of how testing impacted treatment recommendations in node-negative, ER-positive women.[14] The CE study compared treatment recommendations made with RS knowledge (the testing intervention, or "with RS") to those made in a retrospective chart review in which the RS was withheld from the medical oncologist (no testing, or "without RS"). In

this case, the individual-level CE data were made available to us, so we used the individual-level CE data option to construct the virtual population and projected mortality, total costs and quality-adjusted life-years using the same parameters as in Case 1 (Table 2.1). We also performed a PSA using the same standard errors and distributional assumptions as in Case 1.

Tutorial

To facilitate use of the Diagnostics model, we provide a tutorial in the Appendix that shows how to replicate the Case 1 and Case 2 models through the user interface and provides additional technical details and modeling guidance. The tutorial also demonstrates how advanced users can access the R code. This written tutorial is accompanied by a video tutorial to familiarize users with the interface (http://www.fhcrc.org/cantrance). Due to data privacy restrictions, the Case 2 replication in the tutorial uses a public-access approximation of the true individual-level data, generated from the information published in Ademuyiwa et al.[14]

Results

User Interface

The interface for Diagnostics model walks the user through each of the modeling steps outlined in the *CANTRANce Framework* section of the Methods (Figure 2.2). The user specifies modeling choices in a scrolling left panel through a combination of selection boxes, user-defined tables, and numeric entry boxes. Individual-level CE data can be entered using a comma-separated text (.csv) file. Once the CE study data are entered, the left panel updates to reflect those data. The right panel of the interface allows the user to navigate between sections of the model, view help files, and run the model with a viewport to the R session.

Results of Case Studies

The CE studies used as inputs to the Diagnostics model for each case had different population ages and risk distributions (Table 2.2). All Case 1 patients were assumed to be 55, following the Reed model.[30] Average age in Case 2 was also 55, but ages ranged from 29 to 82. About half of patients were classified as low RS in both cases, but the Case 2 data had very few high RS patients compared to Case 1.

Use of the RS decreased adjuvant chemotherapy among low-risk women and increased it among high RS women in both samples. Case 1 data additionally indicated an increase in adjuvant therapy among intermediate RS women with use of the RS (Table 2.2).

The models for both CE studies projected that testing with the RS increases life-years and quality-adjusted life-years (QALYs) at reasonable cost, but with substantial uncertainty (Table 2.3). Benefits of using the RS were greater for the Case 1 models due to the greater proportion of high-risk patients in the population, for whom use of chemotherapy increased from 67% to 100% in Case 1 versus the smaller increase of 88% to 96% in Case 2 (Table 2.2). In addition, more intermediate-risk women were switched onto chemotherapy with the RS in Case 1 (Table 2.2) and this was modeled as beneficial for all of them (Table 2.1). In both the high- and intermediate-risk groups, the survival benefits of chemotherapy outweighed its negative impact on quality of life to result in QALYs gained on average. However, the additional courses of chemotherapy also resulted in greater incremental costs in Case 1. In Case 2, the primary effect of testing was to decrease chemotherapy in low-risk women (Table 2.2). Costs saved by avoiding unnecessary chemotherapy in Case 2 almost completely offset the cost of testing, on average (Table 2.3).

The Case 1 results projected greater incremental life-years and QALYs than those reported by Reed, but in our sensitivity analysis we were able to closely replicate the Reed results (Table 2.3). When we matched the Reed model's life table and assumption limiting recurrence to the first 10 years, we projected 0.23 incremental life-years, 0.18 incremental QALYs, $1,920 incremental total costs per person for with RS versus without RS, and an

incremental cost-effectiveness ratio (ICER) per QALY of $10,772. Reed projected 0.19 incremental life-years, 0.16 incremental QALYs, $1,741 incremental total costs, and an ICER per QALY of $10,881.[30]. The remaining differences between our Case 1 sensitivity analysis and the Reed results are due to the differences in model structures. Our microsimulation model uses a continuous-time approach and represents a population in which parameter inputs define distributions (e.g., exponentially-distributed survival). Their Markov model uses 6-month cycles and represents a homogeneous cohort subject to the exact parameter inputs.

Discussion

The Diagnostic Biomarker module of CANTRANce is a powerful new tool for understanding the likely impact of new diagnostic biomarkers to target cancer treatment. As a free, open-source modeling platform, it is a unique public resource to support timely comparative effectiveness research on emerging diagnostic biomarkers. The user-friendly interface allows investigators to extend their CE results of how testing impacts treatment to mortality, quality-adjusted mortality and costs without needing to commission a customized model. This is made possible by the variety of modeling options available to the user within the Diagnostics model. The model can flexibly accommodate CE data in multiple forms as well as conform to various other user needs, such as the use of disease recurrence as a second intermediate endpoint. The user can also use the interface to easily investigate how altering model assumptions changes results. The availability of all these features in one model holds great potential for quick projections as CE studies of how emerging diagnostic biomarkers impact treatment become available. Head-to-head comparisons of competing diagnostic biomarkers may reveal expected differences—or a lack thereof—that can be used to prioritize further research.

The two case studies presented in this paper validate the capability of the Diagnostics model to produce projections on par with customized models and provide a novel projection of

the impact of testing with a diagnostic biomarker. The Case 1 model closely replicated the Reed model once all assumptions were matched. The Case 2 model suggested less benefit but also lower costs with testing than Case 1. The different results for the two cases highlight that the projections can be sensitive to the CE study on which they are based. Case 1 projected greater benefit and costs than Case 2 due to the different population risk distributions and the differences in the impact of testing indicated by the two CE studies. The interpretation and application of these projected outcomes must thus be considered in context of the original CE study population.

The microsimulation approach we chose for CANTRANce allows us to model the CE study population with realistic variation. In a Markov model like the Reed model, population characteristics are typically simplified in order to maintain a reasonable number of nodes and branches. Microsimulation allows representation of the variation in characteristics across individuals in a population. For example, in a Markov model, if the probability of being recurrence-free at 10 years is 0.968, each simulated individual is subject to that exact probability of being recurrence-free. In a microsimulation model, 0.968 is used to define a distribution for recurrence-free survival, and individuals are assigned different times to recurrence that, when summarized at the population level, equate to a probability of 0.968 at 10 years. One consequence of modeling population variation in this manner is greater uncertainty in results, as seen in Case 1 uncertainty intervals compared to those reported in Reed et al.[30]

The CANTRANce approach of projecting mortality based on CE studies of intermediate endpoints also has its limitations. As with all models, the quality of the projection will reflect the quality of the input data and the accuracy of the assumptions involved. Study populations, such as the ones used in the two case studies, may be small and not representative of major populations of interest. Survival data used to inform model parameters may not be available from populations that resemble the study population. Net survival in particular is not always reported in publications and often must be approximated. In both case studies, we used

recurrence-free survival from published studies to approximate the net survival distribution for time from diagnosis to recurrence. This may overestimate the incidence of recurrence. We used relative survival for advanced SEER cases to project the net survival following disease recurrence. So long as survival for recurrent cases is reasonably approximated by survival for newly-diagnosed advanced cases, this approximation should be reasonable. Even when available, net disease-specific survival may suffer from informative censoring due to competing events such as other-cause death. Finally, the modeling approach of projecting mortality from intermediate endpoint(s) requires that those endpoints behave as surrogates for mortality. We recognize that endpoints for different cancers may vary in the degree to which they are well-established as surrogates.[32] In our examples, we used disease recurrence as a surrogate for breast cancer mortality, a reasonable approach given the large contribution of recurrent cases to mortality and the low likelihood of cure after recurrence.[33] The model also allows for the inclusion of additional covariates for projecting mortality after the intermediate endpoint. Furthermore, if users extend the model to include quality-of-life and cost data, the quality of those inputs will also impact the accuracy of the results.

CANTRANce also has limitations imposed by its generic form. The simple model structure that makes CANTRANce user-friendly also limits its ability to incorporate more complex features represented in customized models. Examples include the shape of the time to event curve, which is constrained to follow an exponential distribution in CANTRANce, the number of key events represented in the disease process, adverse events and heterogeneity in within treatment groups. In the Diagnostics model, we also do not distinguish between treatment recommendations and final treatment decisions, assuming that the recommendations are generally followed. What is gained by losing more complex features is the ability to quickly re-use the model for similar applications without having to build the model from scratch.

The Diagnostics model is a platform for future studies of the value of diagnostic biomarkers for targeting cancer treatment. New diagnostic biomarkers will continue to diffuse into clinical practice before their impact on mortality can be observed. In CANTRANce, we offer a blueprint and modeling system for the extrapolation process. We hope that this will empower CE investigators to project the consequences of their interventions beyond intermediate endpoints to long-term endpoints that can inform the adoption of diagnostic biomarkers.

References

1. Bates, S. Progress towards personalized medicine. *Drug Discov. Today* **15,** 115–120 (2010).
2. Chung, C. & Christianson, M. Predictive and prognostic biomarkers with therapeutic targets in breast, colorectal, and non-small cell lung cancers: A systemic review of current development, evidence, and recommendation. *J Oncol Pharm Pract* (2013). doi:10.1177/1078155212474047
3. Workman, P. The opportunities and challenges of personalized genome-based molecular therapies for cancer: targets, technologies, and molecular chaperones. *Cancer Chemother. Pharmacol.* **52 Suppl 1,** S45–56 (2003).
4. Ow, T. J., Sandulache, V. C., Skinner, H. D. & Myers, J. N. Integration of cancer genomics with treatment selection: From the genome to predictive biomarkers. *Cancer* (2013). doi:10.1002/cncr.28304
5. Chung, C. & Christianson, M. Predictive and prognostic biomarkers with therapeutic targets in breast, colorectal, and non-small cell lung cancers: A systemic review of current development, evidence, and recommendation. *J Oncol Pharm Pract* **20,** 11–28 (2014).
6. Rizzo, S. *et al.* Prognostic vs predictive molecular biomarkers in colorectal cancer: is KRAS and BRAF wild type status required for anti-EGFR therapy? *Cancer Treat. Rev.* **36 Suppl 3,** S56–61 (2010).
7. Linardou, H. *et al.* Assessment of somatic k-RAS mutations as a mechanism associated with resistance to EGFR-targeted agents: a systematic review and meta-analysis of studies in advanced non-small-cell lung cancer and metastatic colorectal cancer. *Lancet Oncol.* **9,** 962–972 (2008).
8. Wang, S. C., Zhang, L., Hortobagyi, G. N. & Hung, M. C. Targeting HER2: recent developments and future directions for breast cancer patients. *Semin. Oncol.* **28,** 21–29 (2001).
9. Bravatà, V., Cammarata, F. P., Forte, G. I. & Minafra, L. 'Omics' of HER2-positive breast cancer. *OMICS* **17,** 119–129 (2013).
10. Dent, S., Oyan, B., Honig, A., Mano, M. & Howell, S. HER2-targeted therapy in breast cancer: a systematic review of neoadjuvant trials. *Cancer Treat. Rev.* **39,** 622–631 (2013).
11. Azim, H. A. *et al.* Utility of prognostic genomic tests in breast cancer practice: The IMPAKT 2012 Working Group Consensus Statement. *Ann. Oncol.* **24,** 647–654 (2013).
12. Lo, S. S. *et al.* Prospective multicenter study of the impact of the 21-gene recurrence score assay on medical oncologist and patient adjuvant breast cancer treatment selection. *J. Clin. Oncol.* **28,** 1671–1676 (2010).
13. Joh, J. E. *et al.* The effect of Oncotype DX recurrence score on treatment recommendations for patients with estrogen receptor-positive early stage breast cancer and correlation with estimation of recurrence risk by breast cancer specialists. *Oncologist* **16,** 1520–1526 (2011).
14. Ademuyiwa, F. O. *et al.* The effects of oncotype DX recurrence scores on chemotherapy utilization in a multi-institutional breast cancer cohort. *Breast Cancer Res. Treat* **126,** 797–802 (2011).
15. Müller, B. M. *et al.* The EndoPredict Gene-Expression Assay in Clinical Practice - Performance and Impact on Clinical Decisions. *PLoS ONE* **8,** e68252 (2013).
16. Webster, J. *et al.* KRAS testing and epidermal growth factor receptor inhibitor treatment for colorectal cancer in community settings. *Cancer Epidemiol. Biomarkers Prev.* **22,** 91–101 (2013).

17. Torrisi, R. *et al.* Potential impact of the 70-gene signature in the choice of adjuvant systemic treatment for ER positive, HER2 negative tumors: a single institution experience. *Breast* **22,** 419–424 (2013).
18. Glasgow, R. E. *et al.* Comparative effectiveness research in cancer: what has been funded and what knowledge gaps remain? *J. Natl. Cancer Inst.* **105,** 766–773 (2013).
19. Lyman, G. H. & Levine, M. Comparative effectiveness research in oncology: an overview. *J. Clin. Oncol.* **30,** 4181–4184 (2012).
20. Rutter, C. M., Zaslavsky, A. M. & Feuer, E. J. Dynamic microsimulation models for health outcomes: a review. *Med Decis Making* **31,** 10–18 (2011).
21. Siebert, U. *et al.* State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-3. *Med Decis Making* **32,** 690–700 (2012).
22. Briggs, A. H. *et al.* Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-6. *Med Decis Making* **32,** 722–732 (2012).
23. Cronin, K. A. & Feuer, E. J. Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival. *Stat Med* **19,** 1729–1740 (2000).
24. National Cancer Institute. SEER Cancer Statistics Review (CSR) 1975-2011. (2014). at <http://seer.cancer.gov/csr/1975_2011/>
25. National Cancer Institute. Surveillance, Epidemiology and End Results Program. at <www.seer.cancer.gov>
26. The Berkeley Mortality Database. at <http://demog.berkeley.edu/~bmd/>
27. Pinsky, P. F. *et al.* Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial. *Am. J. Epidemiol.* **165,** 874–881 (2007).
28. Krahn, M. & Gafni, A. Discounting in the economic evaluation of health care interventions. *Med Care* **31,** 403–418 (1993).
29. Gold, M. R., Siegel, J. E., Russell, L. B. & Weinstein, M. C. *Cost-Effectiveness in Health and Medicine*. (Oxford University Press, 1996).
30. Reed, S. D., Dinan, M. A., Schulman, K. A. & Lyman, G. H. Cost-effectiveness of the 21-gene recurrence score assay in the context of multifactorial decision making to guide chemotherapy for early-stage breast cancer. *Genet. Med.* (2012). doi:10.1038/gim.2012.119
31. Tang, G. *et al.* Comparison of the prognostic and predictive utilities of the 21-gene Recurrence Score assay and Adjuvant! for women with node-negative, ER-positive breast cancer: results from NSABP B-14 and NSABP B-20. *Breast Cancer Res. Treat.* **127,** 133–142 (2011).
32. Fleming, T. R., Prentice, R. L., Pepe, M. S. & Glidden, D. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Stat Med* **13,** 955–968 (1994).
33. Gill, S. & Sargent, D. End points for adjuvant therapy trials: has the time come to accept disease-free survival as a surrogate end point for overall survival? *Oncologist* **11,** 624–629 (2006).

Tables and Figures

Table 2.1

Diagnostics model parameters for the two case studies, based on Reed et al.[30] Distributional assumptions for PSA (probabilistic sensitivity analysis) also follow from Reed et al: beta distributions for probabilities and utilities, normal distributions for costs and log-relative risks, and Dirichlet for the multinomial parameters describing the CE study covariate summaries (not shown).

Parameter	Mean (SE)
10-year distant recurrence-free survival with hormone therapy[a]	
RS low risk	0.968 (0.016)
RS intermediate risk	0.909 (0.043)
RS high risk	0.605 (0.073)
Hazard ratio of chemotherapy on distant recurrence[b]	
RS low risk	1.31 (0.57)
RS intermediate risk	0.61 (0.56)
RS high risk	0.26 (0.31)
5-year relative survival after distant recurrence	0.264
Health state utilities	
Chemotherapy in the first year	0.48 (0.06)
Hormonal therapy/remission	0.68 (0.06)
Distant recurrence	0.42 (0.06)
Medical costs, 2011 USD[c]	
Costs incurred in the first year	
21-gene recurrence score assay	4075
Chemotherapy in the first year	16,947 (1,655)
Absence from work attributable to chemotherapy	12,686
Monitoring and follow-up during remission[d]	1,108 (61)
Distant recurrence, one-time cost	17,478 (2,444)
Patient time during last year of life with metatstatic breast cancer	3,902

[a] *This is a composite statistic used to approximate the net statistic of time to recurrence*
[b] *Standard error is on the log scale*
[c] *5-year annual costs of $105 for hormonal therapy, included by Reed, were not included*
[d] *These costs were applied until recurrence occurred, whereas Reed et al limited them to 10 years*

Table 2.2

Distribution of patients by RS and percent of patients receiving chemotherapy (CTX) within RS groups (Low, Intermediate, and High), for each case study. Data for Case 1 come from Tang et al[31] and Lo et al (N=89).[12] Data for Case 2 come from Ademuyiwa et al (N=276).[14]

	Case 1			Case 2		
	RS Distribution[a]	CTX Without RS[b]	CTX With RS[b]	RS Distribution[a]	CTX Without RS[b]	CTX With RS[b]
Low	53%	42%	13%	51%	37%	9%
Intermediate	19%	50%	36%	40%	47%	47%
High	29%	67%	100%	9%	88%	96%

[a]Column % across RS groups
[b]Row % within RS group

Table 2.3

Lifetime per-patient outcomes for the without RS scenario, and incremental outcomes comparing with RS to without RS, with 95% uncertainty intervals in parentheses. Reed Model results come from reference 29. Case 1 - Sensitivity replicates Reed including the life table and recurrence assumption, while Case 1 replicates Reed using the standard Diagnostics model life table and recurrence assumption. Case 2 is a novel projection of the mortality impact of testing implied by its impact on treatment in Ademuyiwa et al.[14] QALYs = quality-adjusted life-years.

	Reed Model	Case 1 - Sensitivity	Case 1	Case 2
Life-years *without RS*	14.82 (14.46-15.07)	16.02 (14.19-16.71)	16.98 (14.22-18.15)	17.12 (16.19-17.69)
Incremental life-years *with RS*	0.19 (0.09-0.32)	0.23 (-0.13-1.69)	0.33 (-0.23-2.33)	0.05 (-0.03-0.30)
QALYs *without RS*	9.93 (8.12-11.60)	10.74 (8.70-12.56)	11.34 (8.95-13.43)	11.48 (9.42-13.33)
Incremental QALYs *with RS*	0.16 (0.08-0.28)	0.18 (-0.04-1.19)	0.26 (-0.11-1.77)	0.06 (0.00-0.25)
Total costs *without RS*	24,656 (22,599-26,887)	31,943 (17,238-49,269)	34,172 (19,317-51,659)	33,153 (30,658-35,556)
Incremental total costs *with RS*	1,741 (-85-3,710)	1,939 (-15,757-21,363)	1,920 (-15,943-21,291)	26 (-412-482.68)

Figure 2.1

Schematic of the Diagnostics model. The virtual study population is subjected to two scenarios: all are tested for the diagnostic biomarker, or none are tested. The resulting distribution of tailored versus standard therapy for each scenario is based on a CE study. CANTRANce uses additional data from the literature/other studies to project from the time of treatment to mortality.

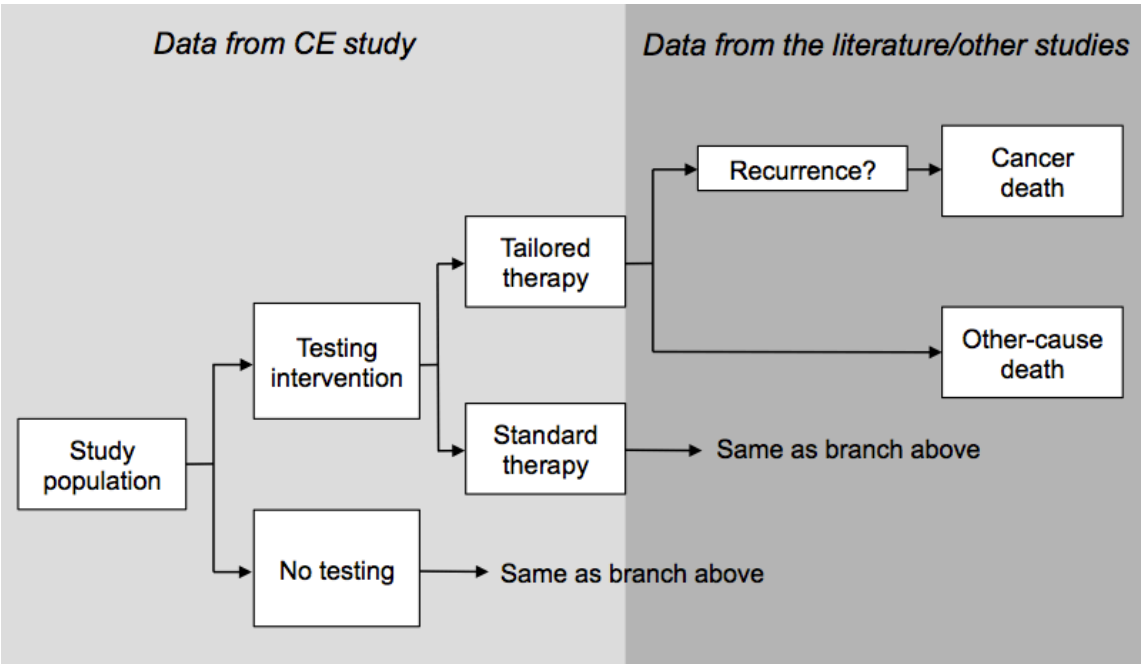
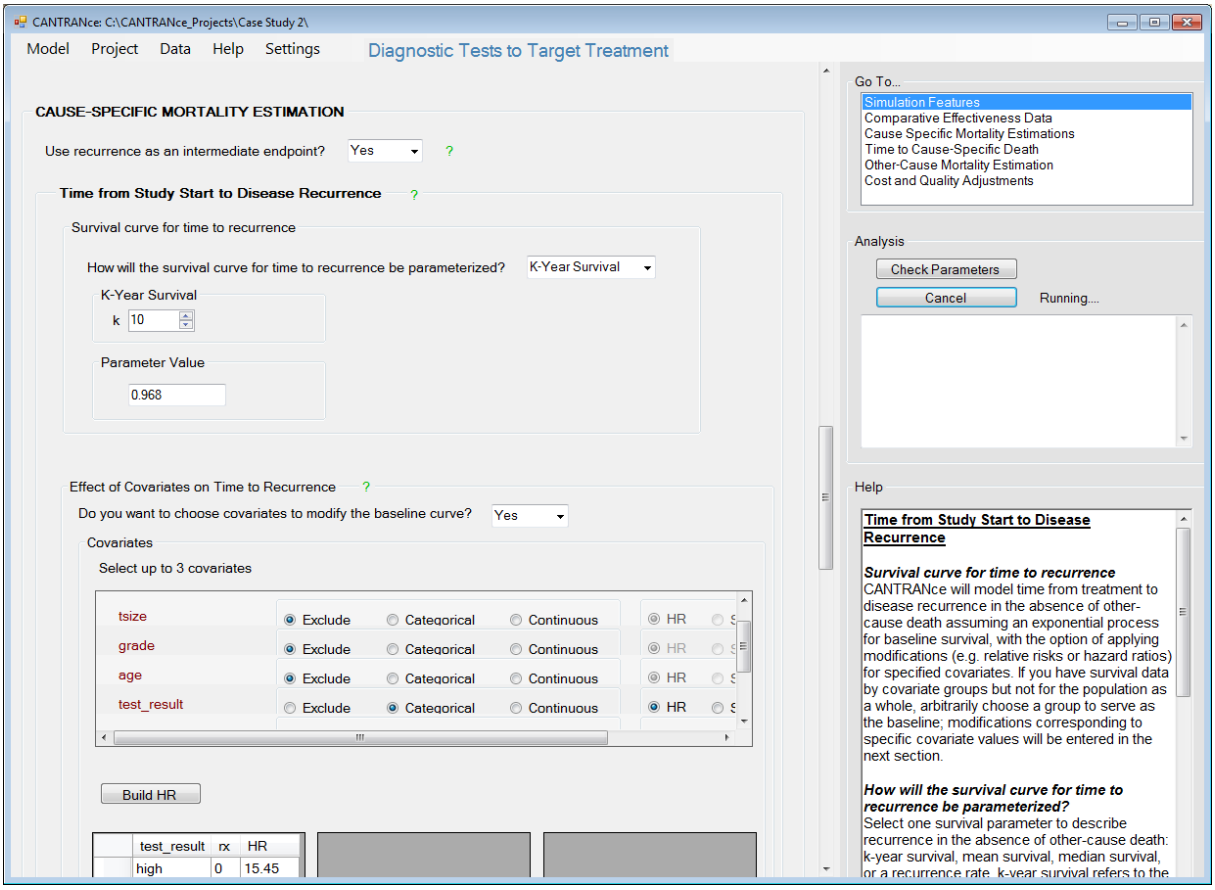


Figure 2.2

Screenshot of the user interface for the Diagnostics model.



Appendix: Tutorial for using the Diagnostics model

Installing the R statistical software

1. Go to http://cran.r-project.org/bin/windows/base and download R for Windows.

2. Open the downloaded .exe file to install R. Use the default installation folder, e.g., C:\Program Files\R\R-3.1.1. Note the location of this folder, as you will need to confirm it later.

***Note**: you will also need an internet connection when you first use the Diagnostics model, to allow the model to auto-download R packages used by the model.*

Installing and initializing the Diagnostics model interface, and accessing R code

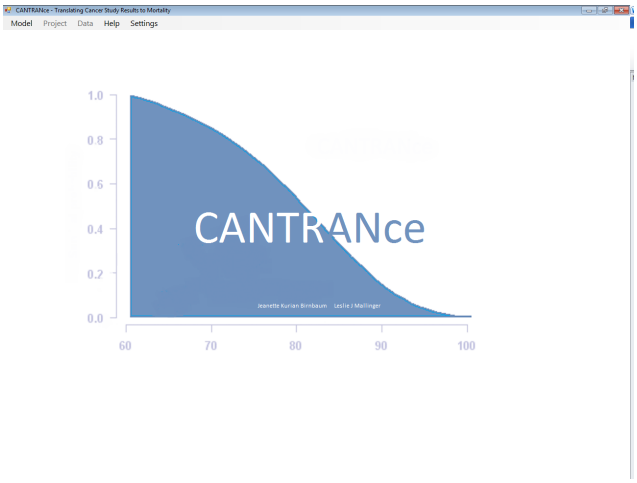
***Note**: This software was built and tested on Windows 7.*

1. Go to http://www.fhcrc.org/cantrance and click on the Windows application link. You may register to receive updates if you wish.

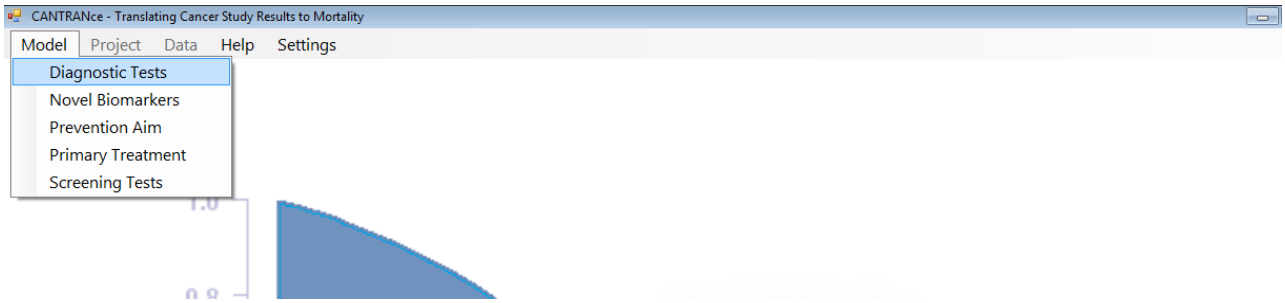
2. Watch the video tutorial, linked from the downloads page (13 min). This will familiarize you with how the interface works and preview for you how the Case 1 model can be run using the interface.

3. Download the setup.exe file.

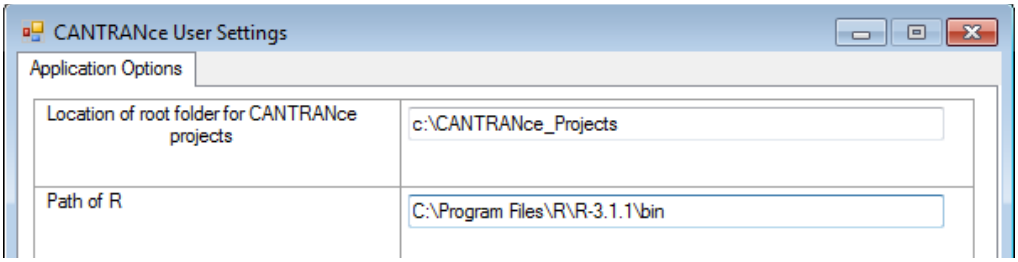
4. Open the .exe file to install CANTRANce. You will see the welcome screen:



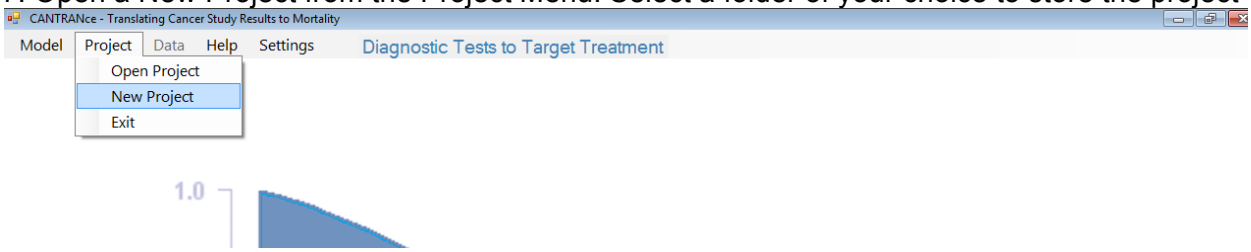
5. Select the Diagnostic Tests model from the Model menu.



6. Go to Settings and edit the “Path of R” to point to the folder where R.exe is stored. If you didn’t choose a special location when installing R, it will most likely be in C:\Program Files\R-3.1.1\bin, as indicated below. You may also edit the “Location of root folder for CANTRANce projects.” Make sure to click “Save” after editing the “Path of R.”



7. Open a New Project from the Project Menu. Select a folder of your choice to store the project



8. Use a file browser window to look at the newly created folder. In the Inputs subfolder, you will see an example individual-level CE data file that will be used in the Case 2 projection, as well as a run_file.R that contains the underlying code for the model.

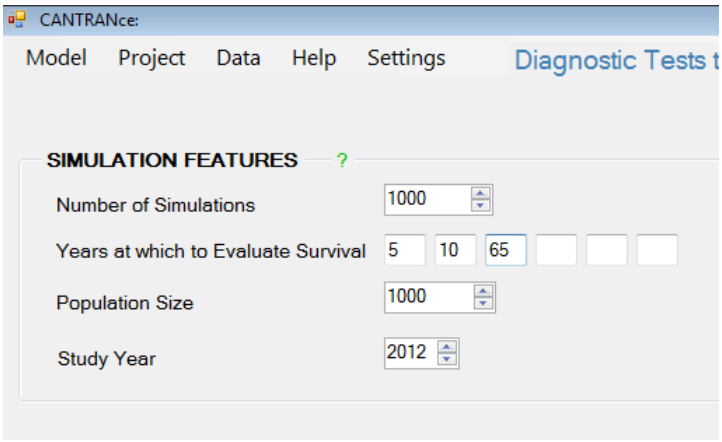
9. To run the Case 1 model, choose “Covariate Proportions” and continue on to the next section, “Running the Case 1 model – Part I.” To run the Case 2 model, choose “Individual-Level Data.” An open-file dialogue box will pop up. Proceed to the section “Running the Case 2 model – Part I” on page 44.

***Note**: Once the interface is initialized, in each section, you may hover your mouse over the green **?** to display a detailed description of each parameter in the “Help” box in the lower right panel. The following instructions for running each case will provide background on each parameter. Read the Help for more details.*

Running the Case 1 model – Part I

***Note**: Running the Case 1 model involves hand-keying in a large table of CE study data. If you want a shorter introduction to the Diagnostics model, skip to the Case 2 model on page 44.*

1. Enter the following parameters in the "Simulation Features" section.

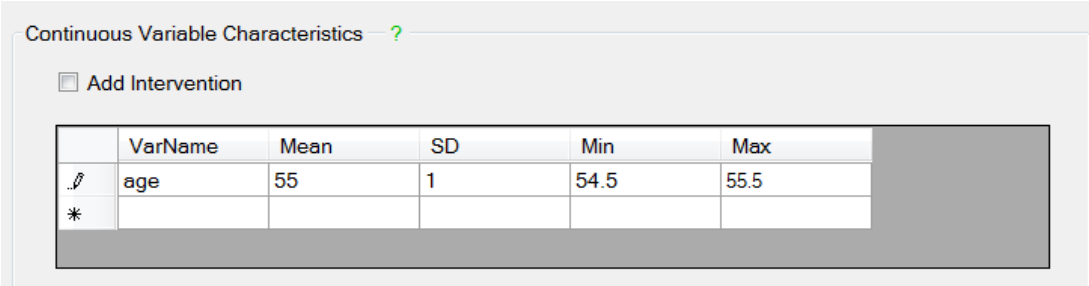


Number of Simulations: 1000 simulations is a standard number of replicates to characterize uncertainty. If you are testing out different models, fewer simulations will produce results faster, but we recommend at least 1000 simulations for a final model. Since the Reed model reported lifetime results, **Years at which to Evaluate Survival** needs a maximum follow-up time that represents lifetime follow-up. Given that all women were aged 55 in the Reed model, we enter 65 years as the maximum follow-up because all women will die by age 120 according to our cohort life table. Because the Reed model presented per-patient rather than population results, the **Population Size** is not very influential and we choose a standard size of 1000. This size impacts the model in two ways: uncertainty due to the simulation process will decrease as (Number of simulations)*(Population Size) increases, and any absolute (versus per-patient) outcomes will refer to this particular population size. [1] We choose 2012 as the **Study Year** given the Reed paper publication date. This parameter will identify birth cohorts for simulated individuals, i.e. women aged 55 in 2012 were born in 1957.

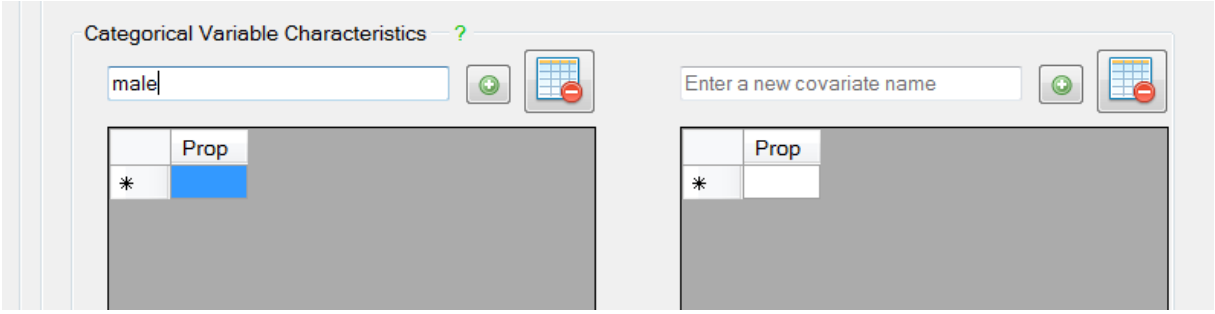
[1] For more on this topic, see Sharif B, Kopec JA, Wong H, Fines P, Sayre EC, Liu RR, Wolfson MC. Uncertainty Analysis in Population-Based Disease Microsimulation Models. Epidemiology Research International. 2012 Jul 25;2012:e610405.

2. Choose “Yes” for **Is this a paired study design?** The Case 1 CE study determined treatment recommendations with and without testing for the *same* patients.

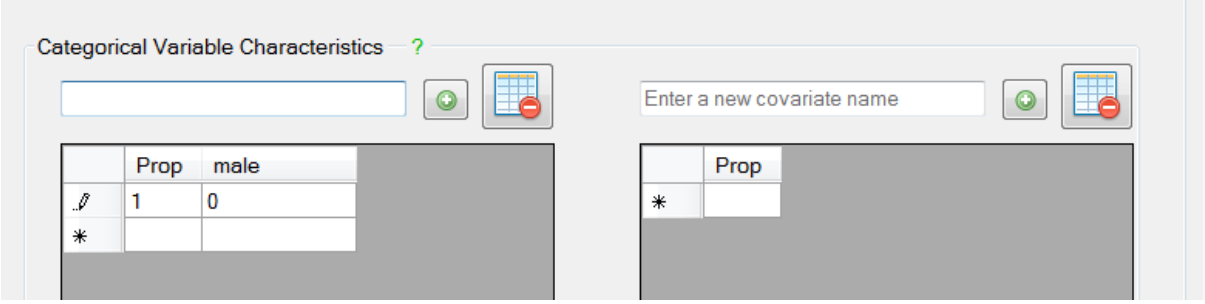
3. Now we will start entering the CE data into the model. Enter “age” as the one continuous variable in **Continuous Variable Characteristics**, as shown below. Since all women were age 55 in the Case 1 model, we will use a truncated normal distribution to normally distribute women’s ages between 54.5 and 55.5 with a mean of 55.0. If the study population had a more complex age distribution, this could be accommodated in several ways – hover over the green ? and read the Help for instructions.



4. Enter “male” as the first categorical variable in **Categorical Variable Characteristics** by typing “male” in as the new covariate name and pressing the green plus button.



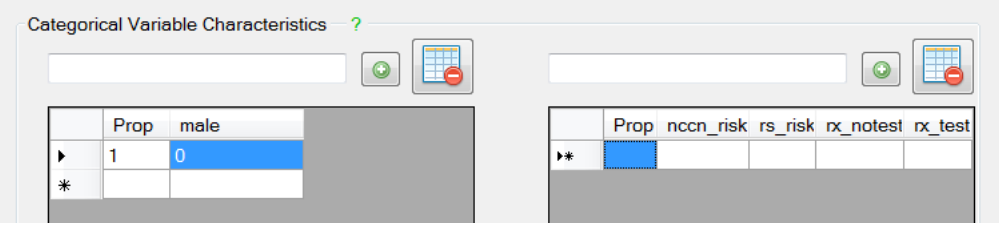
Then enter 0 as the value and Prop(portion) as 1. This tells the model that all individuals are male=0, or female. A population with a mixture of sexes could be modeled by adding a male=1 row and specifying the proportions (e.g., a 50-50 split would be Prop=0.5 for both rows).



5. Enter the remaining CE study data using the 2nd table entry box in **Categorical Variable Characteristics** using the guidelines below. For complete details on how to enter CE study data using **Categorical Variable Characteristics**, please see the Help for this section by hovering over the green ?. It explains how variables can be simulated using either joint or independent distributions. Here, we will focus specifically on the Case 1 CE data entry, in which we need to specify joint distributions for how chemotherapy was recommended according to patients' risk categories.

We will indicate the receipt of chemotherapy (the tailored treatment) in the with-testing scenario using a variable called **rx_test**, and in the without-testing scenario using a variable called **rx_notest**. Because the Case 1 CE study looked at the impact of testing by NCCN and RS risk groups (see Reed Table 1), we must also specify these risk groups using the variables **nccn_risk** and **rs_risk**.

Enter these four variables (**nccn_risk, rs_risk, rx_notest**, and **rx_test**) as the variable names for the 2nd table. *(**Note**: you can resize the column widths to avoid scrolling and use the Tab and Enter keys to navigate within the table. Press the button with the red minus sign to destroy the table and start over.)*



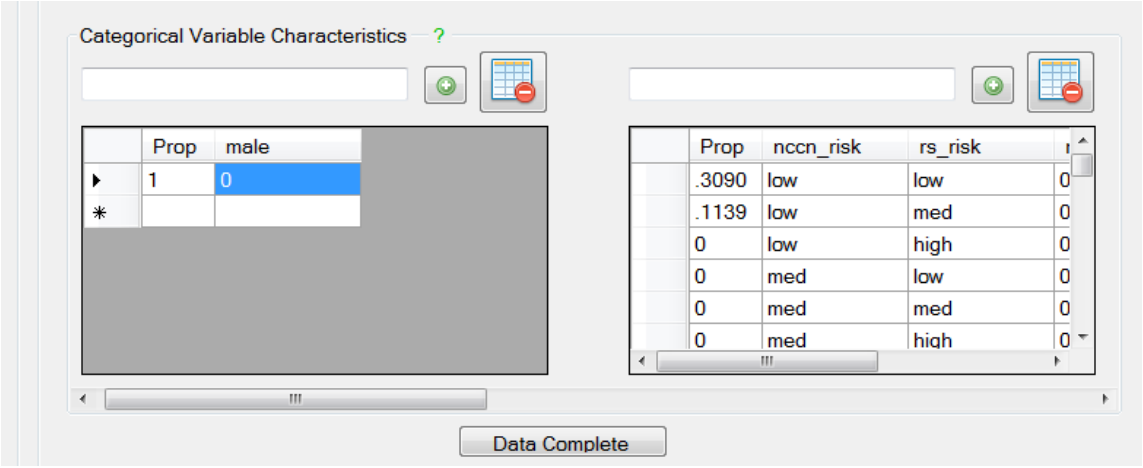
6. Populate this table with 36 rows using the values in the table below. These values are derived from Reed Table 1, by multiplying the proportion of patients in each joint risk group (e.g., NCCN-risk=low and RS-risk=low) by the proportion falling within each of the four possible outcome groups:

1) No chemo in either scenario (rx_notest=0 and rx_test=0)
2) Chemo without testing but no chemo with testing (rx_notest=1 and rx_test=0)
3) No chemo without testing but chemo with testing (rx_notest=0 and rx_test=1)
4) Chemo in both scenarios (rx_notest=1 and rx_test=1)

Prop	nccn_risk	rs_risk	rx_notest	rx_test
0.3090	low	low	0	0
0.1139	low	med	0	0
0	low	high	0	0
0	med	low	0	0
0	med	med	0	0
0	med	high	0	0
0	high	low	0	0
0	high	med	0	0
0	high	high	0	0
0	low	low	1	0
0	low	med	1	0
0	low	high	1	0
0.0640	med	low	1	0
0.0137	med	med	1	0
0	med	high	1	0
0.0729	high	low	1	0
0.0233	high	med	1	0
0	high	high	1	0
0.0145	low	low	0	1
0.0119	low	med	0	1
0.0779	low	high	0	1
0	med	low	0	1
0	med	med	0	1
0	med	high	0	1
0	high	low	0	1
0	high	med	0	1
0	high	high	0	1
0	low	low	1	1
0	low	med	1	1
0	low	high	1	1
0.0213	med	low	1	1
0.0223	med	med	1	1
0.0645	med	high	1	1
0.0243	high	low	1	1

0.0380	high	med	1	1
0.1285	high	high	1	1

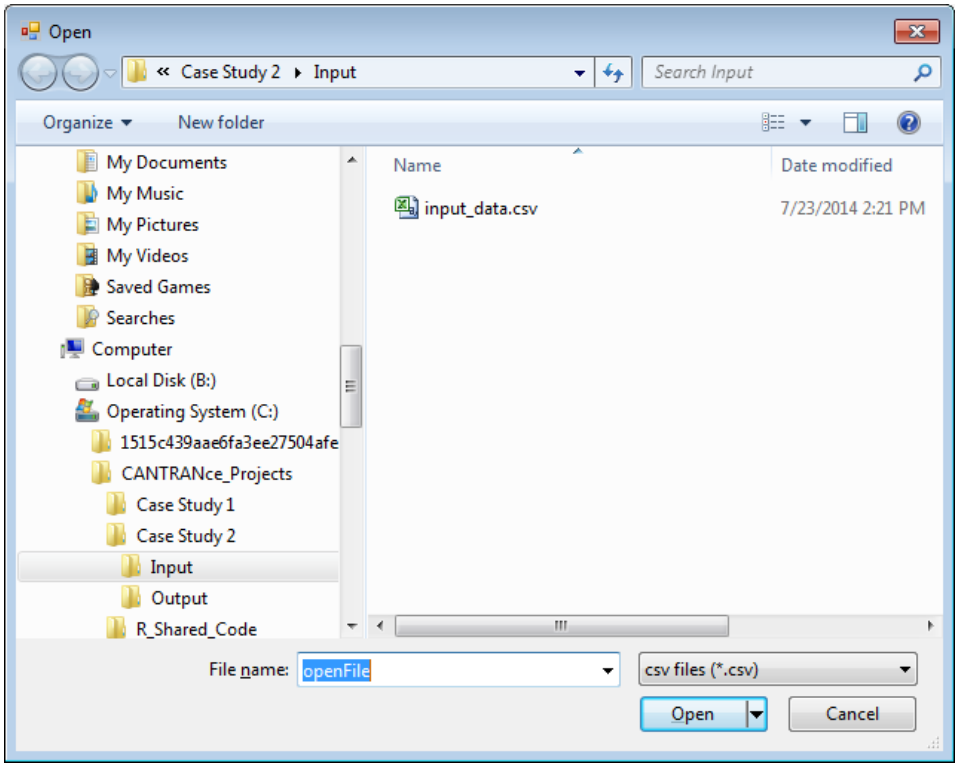
7. When you are done filling out the table, press the “Data Complete” button under the table.



8. To continue running the Case 1 model, proceed to the section “Running the Case 1 or Case 2 model – Part II” on page 47.

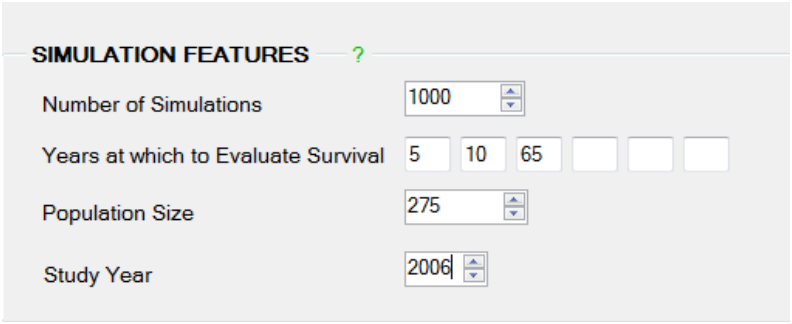
Running the Case 2 model – Part I

1. When you choose “Individual-level Data” as your option from the Data menu, it will open a browser where you can choose a .csv file. For the Case 2 model, use the “input_data.csv” file that is located in the Input folder of the new project folder you created.



This file is a public-access approximation to the the Ademuyiwa CE study data used in Case 2 (reference 14 in the manuscript). More details are given on the next page.

2. Enter the following parameters in the “Simulation Features” section.



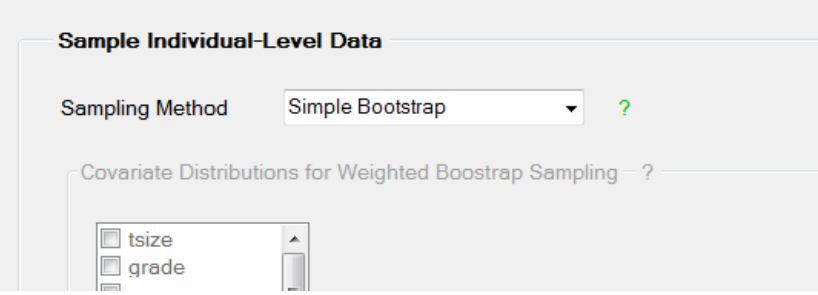
Number of Simulations: 1000 simulations is a standard number of replicates to characterize uncertainty. Since we want to estimate lifetime results, **Years at which to Evaluate Survival** needs a maximum follow-up time that represents lifetime follow-up. Given that the average age was 55 in the Ademuyiwa CE study, we enter 65 years as the maximum follow-up because all women will die by age 120 according to our cohort life table. **Population Size** is 275, the number of subjects in the CE study, and **Study Year** is 2006, the year the CE data were collected.

3. The **Comparative Effectiveness Data** section is grayed out because all the data are contained in the input_data.csv file. The first few rows of the file are shown below.

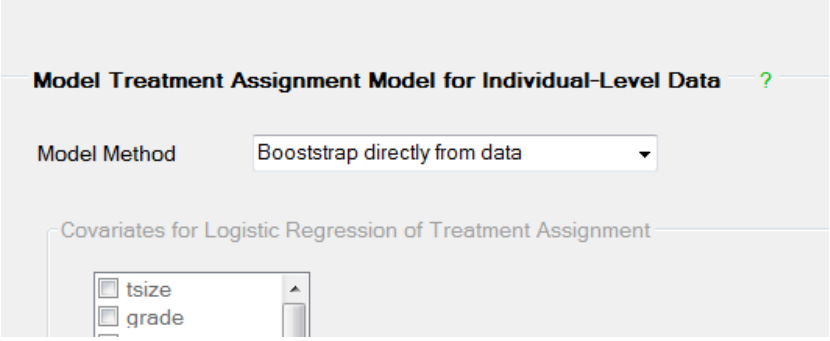
	A	B	C	D	E	F	G	H
1	tsize	grade	age	test_result	rx_notest	rx_test	male	
2	>2cm	1	40	low	1	0	0	
3	1.1-2cm	1	60	low	1	0	0	
4	<1cm	1	45	low	1	0	0	
5	>2cm	1	60	low	1	0	0	
6	1.1-2cm	1	65	low	1	0	0	
7	1.1-2cm	1	50	low	1	0	0	
8	<1cm	1	60	low	1	0	0	
9	1.1-2cm	1	65	low	1	1	0	

Each record represents a *simulated version of* one of the patients in the Ademuyiwa CE study data used in Case 2. All patients were female, so male=0 for all records. The individual values of the remaining covariates do not exactly match the true CE study data, but the population-level statistics for **age**, **test_result** (21-gene recurrence score risk group), **rx_notest** (treatment recommendation without testing, 1=chemo+hormone 0=hormone), and **rx_test** (treatment recommendation with testing) match those published in the Ademuyiwa et al paper, Tables 4 and 5 (see reference 14). The **tsize** and **grade** covariates are fake covariates you can play around with.

4. Leave **Sample Individual-Level Data** as “Simple Bootstrap.” This means that when the data is resampled for each simulation to represent sampling uncertainty, no weights will be used. Alternatively, a weighted bootstrap could be used to model a hypothetical population with different covariate distributions (see the Methods section of the manuscript). For example, in the Case 2 data the 21-gene recurrence risk distribution is 51% low-risk, 40% medium-risk and 9% high-risk (see Table 2 in the manuscript). If you wanted to model outcomes for a higher-risk population, you could choose “Weighted Bootstrap” below, select “test_result” as the variable, and enter different proportions for the risk categories, e.g. 40% low-risk, 40% medium-risk, and 20% high-risk. The model would then resample the data, oversampling higher-risk cases and undersampling low-risk cases to match the distribution you specified. The remaining projection would pertain to the altered population.



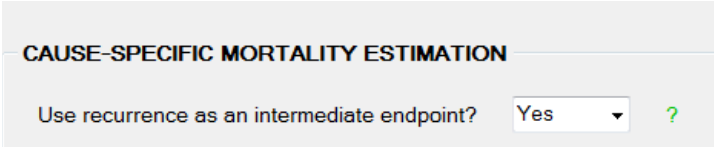
5. Leave **Model Treatment Assignment for Individual-Level Data** as “Bootstrap directly from data.” This means that the model will take the values for treatment recommended in each scenario directly from the CE study data, rather than prompting the user for a logistic regression model from which to simulate treatment recommendations (see the Methods section of the manuscript for reasons why you might choose to do a logistic regression rather than bootstrap directly from data).



6. Continue to “Running the Case 1 or Case 2 model – Part II” on the next page.

Running the Case 1 or Case 2 model – Part II

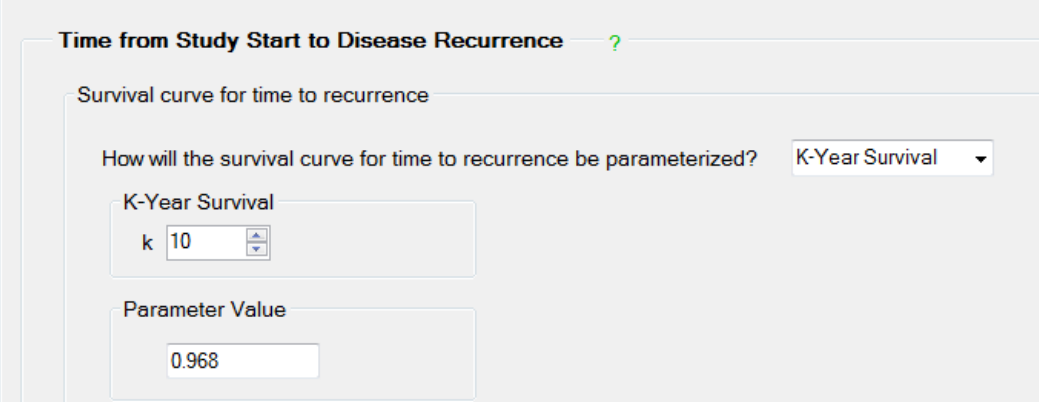
1. Choose “Yes” to **Use recurrence as an intermediate endpoint?**



In both models, we use recurrence as a second intermediate endpoint. This reflects the available long-term data on the particular diagnostic test intervention under study, the 21-gene recurrence score (RS). For this test, time from treatment to recurrence has been reported for risk subgroups defined by the test, but time from treatment to mortality has not. If time to mortality data were available, we would not have to use recurrence as an intermediate endpoint.

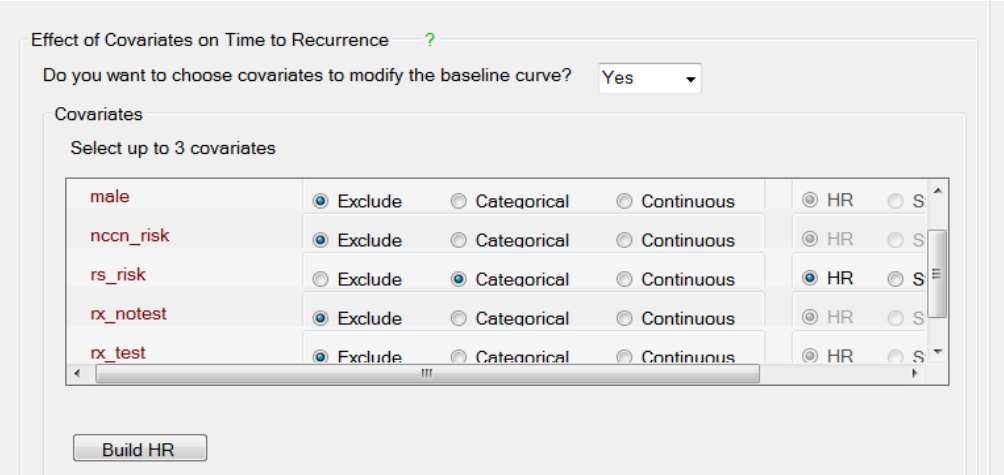
2. We will now enter the time to recurrence parameters from Table 1 of the manuscript.

Enter the parameters for **Survival curve for time to recurrence** as shown below. We choose the 10-year distant recurrence statistic for RS low risk women not receiving chemotherapy to initially define the time to recurrence survival curve. The curve is specified here defines the baseline curve to which subsequent hazard ratios will refer.

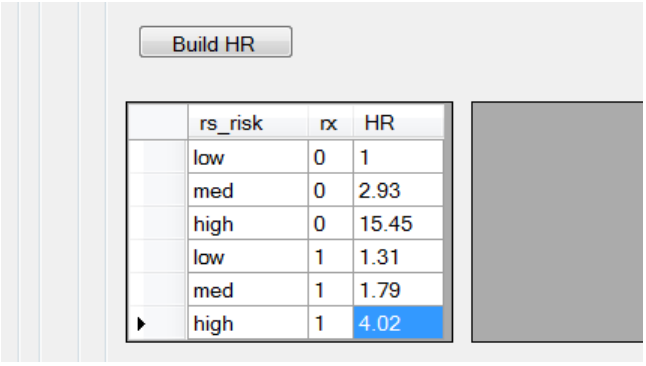


Note that this specification of a 10-year survival of 0.968 is particular to this specific application: it represents time to breast cancer recurrence for RS low risk women. In other words, 10 years after treatment, we expect 96.8% of RS low risk women who did not receive chemotherapy to be recurrence-free. The appropriate survival statistic will change if the model is applied to a different diagnostic test and/or different cancer.

3. Now we will specify different times to recurrence for different RS risk and treatment groups. Choose “Yes” to **Do you want to choose covariates to modify the baseline curve?** Then, select “Categorical” and “HR” for the covariate “rs_risk” (Case 1) or “test_result” (Case 2), and click “Build HR.” We do this because Table 1 contains data on how time to recurrence varies by RS risk group and by treatment. While we primarily need to know how time to recurrence varies for tailored versus standard treatment, having these data stratified by RS risk group can additionally help us be more precise in our projection. We will specify this variation in the form of hazard ratios (HRs).



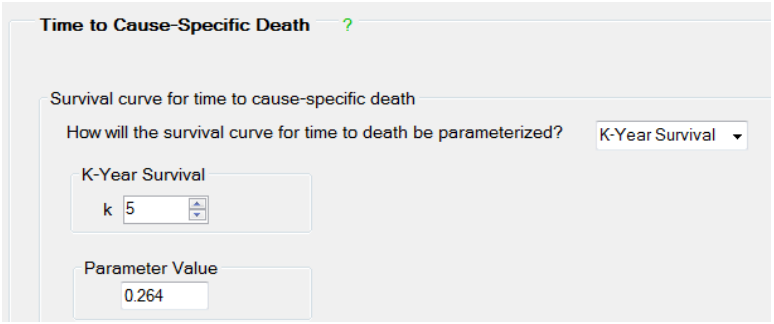
4. Populate the table that is created as shown below. These HRs are derived from Table 1 in the paper.* The baseline group of RS low risk, no chemotherapy (rx=0) is indicating using HR=1. This is the group that we defined above as having a 10-year time to recurrence parameter of 0.968. All other hazard ratios are relative to this group. *Note: the covariate “rs_risk” shown below is called “test_result” in the Case 2 data, and the order of rows may be different. Please make sure you do not simply enter the HRs in order, but check that you are entering the correct HR for the risk and treatment group indicated in each row.*



	rs_risk	rx	HR
	low	0	1
	med	0	2.93
	high	0	15.45
	low	1	1.31
	med	1	1.79
▶	high	1	4.02

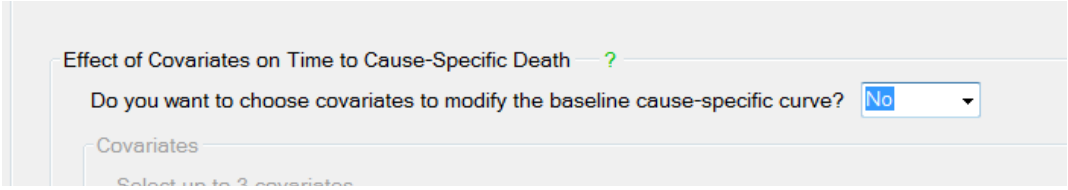
* Hazard ratios are the ratios of rates. We enter HRs here relative to the low-risk, no chemotherapy group, so the HRs are the ratios of each group’s recurrence rate to that group. The Table 1 parameters are on the survival scale, so to compute the HRs we must transform back to the rate scale using rate = log(survival)/k, where k=the year at which survival was evaluated. For example, to compute the HR for med-risk, no chemo to low-risk, no chemo, we have the ratio of rates [log(0.909)/10]/[log(0.968)/10] = 2.93. To get the HR for med-risk with chemo relative to low-risk, no chemo, we multiple the med-risk, chemo versus med-risk, no chemo with the HR we just computed, 0.61*2.93 = 1.79.

5. Enter the parameters for **Survival curve for time to cause-specific death** as shown below. Because we chose to go use disease recurrence as a second intermediate endpoint, this survival curve refers to the time from recurrence to mortality. If we had not chosen recurrence as a second intermediate endpoint, this time would reflect time from treatment to mortality.

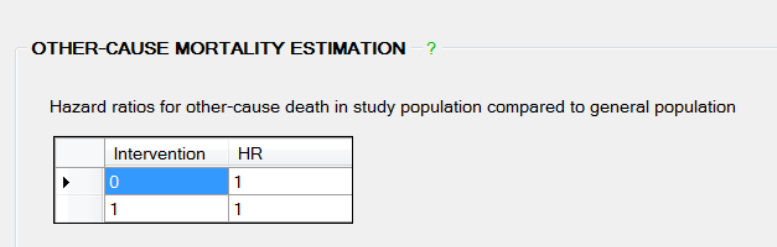


As indicated in Table 1 of the manuscript, this 5-year survival of 0.264 is the relative breast cancer survival for women after distant recurrence of breast cancer. In other words, 5 years after distant recurrence, we expect 26.4% of women to have survived breast cancer death *net of death from other causes* (please see Methods and Discussion of manuscript regarding net survival).

6. Choose "No" to **Do you want to choose covariates to modify the baseline cause-specific curve?** We do this because we assume that once patients recur, their remaining cause-specific survival does not vary significantly by any covariates. If we had evidence and data that available covariates significantly modified time from recurrence to mortality, we could choose "Yes" and enter hazard ratios for those covariates here.



7. Do not modify the **Other-Cause Mortality Estimation** section. We choose to make no assumptions about how other-cause (non-breast-cancer) mortality in our study population may differ from that of the general US population, which means that the hazard ratios (HRs) stay as 1. However, it is plausible that a study population may have a different overall survival profile from the US population. Trial populations, for example, may be healthier due to the selection criteria. Observational screening cohorts may be healthier due to the "healthy screenee effect" in which individuals who select to screen are a healthy subset of the general population. While it may be difficult to precisely quantify these types of effects, a HR lower than 1 could be used to explore the potential impact of the intervention on a healthier-than-US population.

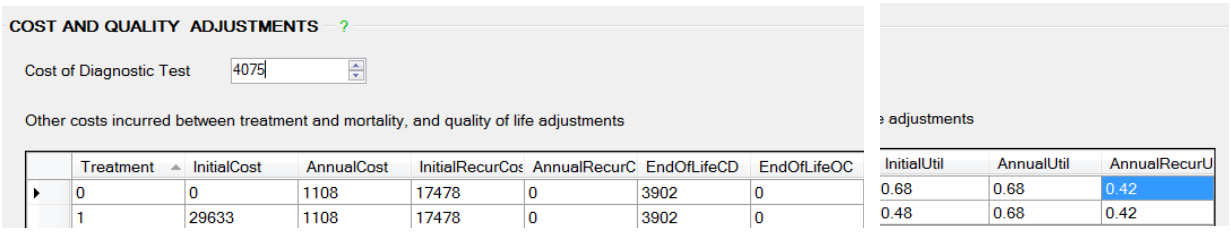


A NOTE ABOUT MODELING TIMES TO EVENT

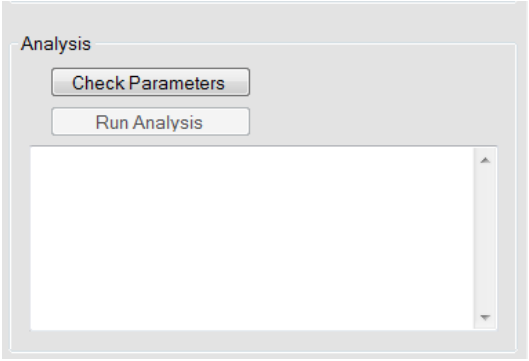
As noted in the manuscript, CANTRANce models times to event using exponential survival curves. These curves assume a constant rate of event (e.g., recurrence or mortality), which is why you only had to enter one statistic to define the time to recurrence or time to mortality curve. While the constant rate assumption is convenient because it only requires one parameter, it may not always well-represent the true survival pattern. That being said, in many applications it will approximate true survival closely enough to produce informative results.

For a more extensive discussion of ways to represent survival, please see:
Connock M, Hyde C, Moore D. Cautions regarding the fitting and interpretation of survival curves: examples from NICE single technology appraisals of drugs for cancer. Pharmacoeconomics. 2011 Oct;29(10):827–837. PMID: 21770482

8. Enter the following parameters in the **Cost and Quality Adjustments** section. Treatment=0 refers to patients who do not get chemotherapy (standard treatment), and Treatment=1 to those who do (tailored treatment). For Treatment=1, **InitialCost** is the sum of chemotherapy costs in the 1st year plus absence of work attributable to chemotherapy. All other costs follow directly from Table 1 of the manuscript. See the Help (hover over the green ?) for a detailed description of each column of the input table.



In the right panel, click “Check Parameters.” If no warning messages appear, you have entered all inputs without errors. Correct any errors until clicking “Check Parameters” results in no messages. Then, click “Run Analysis.”



The model should take under 10 minutes to run for Case 1 and under 5 minutes for Case 2.

Interpreting Results

This section will walk the user through interpreting the results from both Cases. Identically-named .csv and .html files contain the same data.

Outputs that validate Inputs

File: treatment_summary.csv/html

Mean and quantiles of treatment probability cross-classifications between scenarios (Group 0 = without diagnostic test; Group 1 = with diagnostic test)

Case 1

Group 0 Treatment	Group 1 Treatment	Mean	2.5%	50%	97.5%
0	0	0.423	0.392	0.423	0.455
0	1	0.104	0.086	0.104	0.123
1	0	0.173	0.151	0.173	0.199
1	1	0.299	0.271	0.299	0.329

Table 2 Treatment probabilities cross-classified by RS-guided and non-RS-guided strategies

	RS-guided strategy		
Non-RS-guided strategy	**Hormonal therapy**	**Chemo-therapy**	**Total**
Hormonal therapy, %	42.2	10.5	52.7
Chemotherapy, %	17.4	29.9	47.3
Total, %	59.6	40.4	100.0

RS, recurrence score.

Case 2

Group 0 Treatment	Group 1 Treatment	Mean	2.5%	50%	97.5%
0	0	0.430	0.374	0.429	0.487
0	1	0.123	0.084	0.124	0.164
1	0	0.266	0.215	0.265	0.316
1	1	0.181	0.138	0.182	0.229

Table 4 ODX-blinded recommendation and actual treatment received based on ODX score

	CTX received				All	
	No		Yes			
	N	%	*N*	%	*N*	%
ODX-blinded CTX recommendation						
No	117	42.3	34	**12.3**	151	54.7
Yes	71	**25.7**	54	19.7	125	45.3
All	188	68.0	88	32.0	276	100

The treatment_summary table shows the proportion of the virtual population falling into each of the four possible groups of treatment outcomes:

1) No chemo in either scenario (Group 0 Treatment=0 and Group 1 Treatment=0)
2) No chemo without testing but chemo with testing (Group 0 Treatment=0 and Group 1 Treatment=1)
3) Chemo without testing but no chemo with testing (Group 0 Treatment=1 and Group 1 Treatment=0)
4) Chemo in both scenarios (Group 0 Treatment =1 and Group 1 Treatment=1)

The 95% uncertainty intervals (2.5^{th} percentile – 97.5^{th} percentile) arise from the distribution across the 1000 simulations. An important feature of the model to understand is that each run of the model uses one parameter set, the parameters you entered, but that one parameter set is used for multiple simulations. As described in the Discussion, each simulation reflects the same population-level parameters, but the individual values of covariates are different for the same virtual person across simulations. The results are presented as a summary across simulations: the mean across simulations is the point estimate, and the 2.5% and 97.5% percentiles are used to construct a 95% uncertainty interval. We also sometimes present the median, the 50% percentile.

In Case 1, you can see that these proportions closely approximate those given in Reed Table 2 (RS-guided = Group 1 and Non-RS-guided = Group 0). For example, we simulate 0.423 or 42.3% receiving treatment=0, hormonal therapy, in both scenarios, and below, Reed shows 42.2%. We simulate 0.299 or 29.9% receiving chemo in both scenarios (treatment=1), exactly as Reed does. Similarly, the virtual population in Case 2 closely approximates Ademuyiwa Table 4 (CTX received = Group 1 and ODX-blinded recommendation = Group 0). For example, both indicate 12.3% of patients were switched onto chemotherapy with testing.

File: net_recurrence_free_survival.csv/html

10-year recurrence-free survival in the simulated population in the absence of other-cause death

Case 1

Covariate	Value	Treatment	K-Time Survival	2.5%	50%	97.5%
All	All	All	0.927	0.920	0.927	0.934
All	All	0	0.937	0.928	0.936	0.945
All	All	1	0.914	0.903	0.914	0.925
rs_risk	low	0	0.968	0.960	0.968	0.975
rs_risk	med	0	0.909	0.889	0.909	0.929
rs_risk	high	0	0.605	0.515	0.605	0.681
rs_risk	low	1	0.958	0.942	0.958	0.973
rs_risk	med	1	0.943	0.921	0.943	0.965
rs_risk	high	1	0.878	0.859	0.878	0.896

Case 2

Covariate	Value	Treatment	K-Time Survival	2.5%	50%	97.5%
All	All	All	0.940	0.926	0.940	0.952
All	All	0	0.943	0.928	0.943	0.958
All	All	1	0.935	0.913	0.936	0.956
test_result	high	0	0.581	0.139	0.600	0.920
test_result	low	0	0.968	0.953	0.968	0.982
test_result	med	0	0.909	0.873	0.910	0.940
test_result	high	1	0.875	0.796	0.878	0.939
test_result	low	1	0.958	0.917	0.960	0.991
test_result	med	1	0.943	0.914	0.944	0.967

The net_recurrence_free_survival table calculates the rate of recurrence in the virtual population. The table uses the same statistic type inputted by the user, and reports recurrence in the whole population ("All", "All", "All" row), by treatment status ("All", "All", 0/1 rows), and by the groups by which the user inputted hazard ratios (HRs) or different survival statistics (remaining rows). In both case studies, we put in a baseline 10-year survival statistic and hazard ratios by the six RS-treatment groups:

1) low, no chemo (treatment=0)
2) med, no chemo
3) high, no chemo
4) low, chemo (treatment=1)
5) med, chemo
6) high, chemo

The table thus reports 10-year survival statistics for recurrence for each of these groups. You can see that the low-0 row (RS low-risk, no chemo) has a statistic of 0.968, the baseline survival we put in. The other statistics must be converted back to HRs to be compared to our inputs. This must be done on the log scale, e.g. the HR for med-0 (RS med-risk, no chemo) in Case 1 is log(0.909)/log(0.968) = 2.93, exactly what we put in as the HR.

File: net_cause_specific_survival.csv/html
5-year cause-specific survival in the simulated population in the absence of other-cause death

Case 1

Covariate	Value	Treatment	K-Time Survival	2.5%	50%	97.5%
All	All	All	0.263	0.226	0.263	0.302
All	All	0	0.263	0.206	0.263	0.315
All	All	1	0.263	0.208	0.263	0.317

Case 2

Covariate	Value	Treatment	K-Time Survival	2.5%	50%	97.5%
All	All	All	0.263	0.178	0.265	0.349
All	All	0	0.262	0.158	0.263	0.372
All	All	1	0.264	0.136	0.265	0.402

The net_cause_specific_survival table calculates the rate of cause-specific death in the virtual population. When recurrence is used as an intermediate endpoint, as in both case studies, this refers to the time from recurrence to cause-specific death. The format of this table is similar to that of the previous table. The type of statistic reported again corresponds to that input by the user. Since we inputted 5-year survivals, that is the output we see, and the results approximate the value of 0.264 we put in.

Outputs that present results

File: survival_summary.csv

Mean, median, and k-year all-cause, crude, and net survival by testing scenario (Group 0 = without diagnostic test; Group 1 = with diagnostic test)

Case 1

Measure	Group	Mean Survival	Lower	Upper	Median Survival	Lower	Upper	5-Year Survival	Lower	Upper	10-Year Survival	Lower	Upper	65-Year Survival	Lower	Upper
All-Cause Survival	0	25.2	24.5	25.9	26.0	24.9	27.1	0.948	0.934	0.961	0.865	0.843	0.886	0.000	0.000	0.000
All-Cause Survival	1	25.8	25.0	26.5	26.7	25.6	27.8	0.954	0.940	0.966	0.879	0.858	0.900	0.000	0.000	0.000
Net Survival	0	192.6	178.3	208.0	107.4	95.8	119.2	0.978	0.969	0.987	0.940	0.925	0.954	0.634	0.604	0.664
Net Survival	1	203.9	188.7	220.4	115.8	104.5	127.3	0.984	0.976	0.992	0.955	0.942	0.967	0.668	0.639	0.698
Crude Survival	0							0.978	0.970	0.987	0.942	0.928	0.956	0.818	0.793	0.841
Crude Survival	1							0.984	0.976	0.992	0.957	0.944	0.969	0.848	0.825	0.870

Case 2

Measure	Group	Mean Survival	Lower	Upper	Median Survival	Lower	Upper	5-Year Survival	Lower	Upper	10-Year Survival	Lower	Upper	65-Year Survival	Lower	Upper
All-Cause Survival	0	23.7	22.0	25.3	22.3	20.1	24.5	0.922	0.889	0.953	0.817	0.770	0.860	0.003	0.000	0.008
All-Cause Survival	1	23.7	22.0	25.4	22.3	20.1	24.6	0.922	0.889	0.953	0.817	0.768	0.860	0.003	0.000	0.008
Net Survival	0	202.2	176.9	230.8	123.8	103.2	147.1	0.984	0.969	1.000	0.957	0.931	0.979	0.687	0.634	0.742
Net Survival	1	213.0	184.8	242.8	128.0	105.6	153.3	0.984	0.968	0.995	0.957	0.933	0.980	0.694	0.639	0.746
Crude Survival	0							0.985	0.970	0.997	0.961	0.937	0.981	0.874	0.833	0.909
Crude Survival	1							0.985	0.970	0.997	0.961	0.937	0.982	0.877	0.836	0.913

This table presents three types of survival statistics (“Measure” column): all-cause, net and crude survival. All-cause survival is overall survival, i.e. survival from all causes of death. Net and crude survival are different ways of measuring cause-specific death. Net survival censors other causes of death, effectively creating a hypothetical world in which everyone dies of the cancer eventually. Crude survival removes those who have died from other causes of death from the at-risk population, so survival plateaus once everyone has died of other causes. Mean and median crude survivals do not exist for the above projections because survival plateaus before even 50% of the population dies. Crude survival can be thought of as the inverse of cumulative incidence.

interface. The “Lower” and “Upper” columns represent the 95% uncertainty interval across simulations.

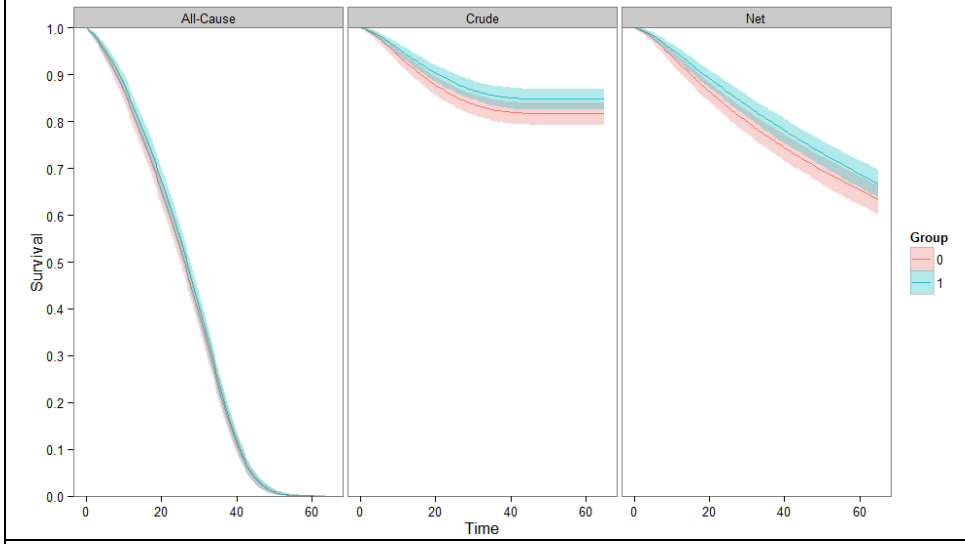
Some highlights of these results: for both the 5- and 10-year Case 1 survivals, the model projects that testing saves 1 life per every 100 patients (1% better survival in the with-testing group). The 65-year survival is a lifetime estimate that suggests that over the lifetime of the whole Case 1 population, about 3 breast cancer deaths per 100 cases are averted by testing. In the Case 2 model, survival is virtually identical for the with- and without-testing groups by all measures, except for long-term net survival.

These numeric summaries are represented graphically in the survival_summary.png file, shown below.

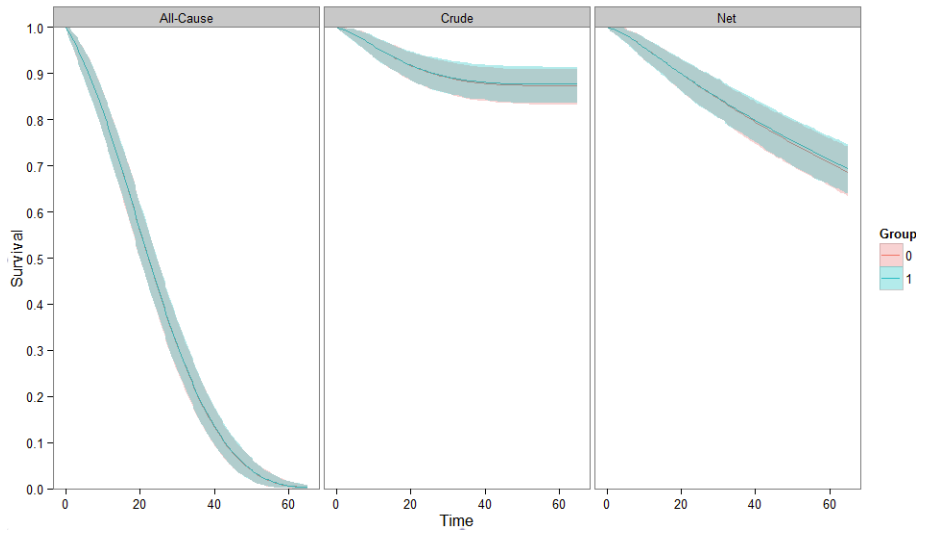
File: survival_summary.png

All-cause, crude and net survival from time of treatment (Time=0) to the maximum follow-up time specified. The line plotted is the median across simulations and the uncertainty bounds represent the 2.5% and 97.5% quantiles.

Case 1



Case 2



The graphs allow us to quickly see the greater advantage of testing projected by Case 1 than by Case 2, in which the Group 1 and Group 0 curves overlap almost entirely. There is also greater uncertainty across simulations in Case 2, due to the smaller population size (N=275) compared to Case 1 (N=1000).

File: person_years_saved.csv/html

Total life-years saved by testing (non-discounted)

Case 1

Mean	585
Lower Bound	7
Upper Bound	1179

Case 2

Mean	15
Lower Bound	-262
Upper Bound	268

This table gives the non-discounted, total number of life-years saved by testing over the maximum follow-up time, which we set as 65 years. This is a population-level statistic that will reflect the size of the population, i.e. larger populations will have more potential for more absolute life-years saved.

File: ICERS.csv/html

Per-person life-years, quality-adjusted life-years (QALYs), costs, and cost-effectiveness ratios (discounted at 3%)

Case 1

	Group 0 Life-Years	Group 1 Life-Years	Group 0 QALYs	Group 1 QALYs	Group 0 Costs	Group 1 Costs	Incremental Life-Years	Incremental QALYs	Incremental Cost	ICER for Life-Years	ICER for QALYs
Mean	16.96	17.28	11.32	11.58	34106	36131	0.32	0.25	2025	6361	8008
Lower	16.58	16.86	11.06	11.29	33103	35139	0.01	0.03	1036	73140	29843
Median	16.97	17.29	11.33	11.58	34105	36127	0.32	0.25	2016	6315	8018
Upper	17.32	17.67	11.58	11.85	35082	37153	0.63	0.48	2999	4763	6311

Case 2

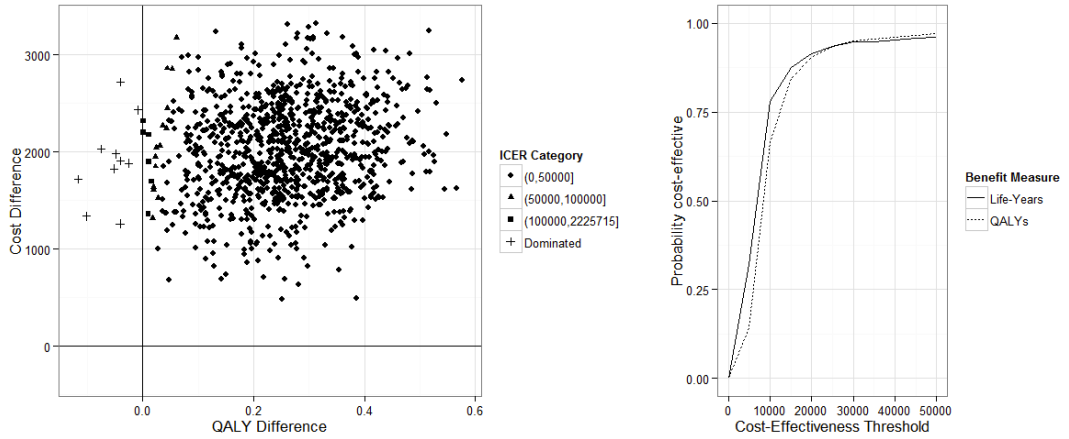
	Group 0 Life-Years	Group 1 Life-Years	Group 0 QALYs	Group 1 QALYs	Group 0 Costs	Group 1 Costs	Incremental Life-Years	Incremental QALYs	Incremental Cost	ICER for Life-Years	ICER for QALYs
Mean	15.90	15.92	10.64	10.69	31422	31246	0.02	0.04	-176		
Lower	15.05	15.05	10.05	10.10	29586	29346	-0.47	-0.31	-2237		
Median	15.90	15.92	10.65	10.69	31411	31230	0.02	0.05	-188		
Upper	16.74	16.77	11.22	11.27	33426	33054	0.51	0.39	1839	3633	4721

These results approximate the Case 1 and Case 2 results reported in the manuscript Table 3. For Case 1, the small differences arise from the truncated decimals used when hand-keying in the table of proportions. For Case 2, the differences arise from using a simulated version of the true CE data.

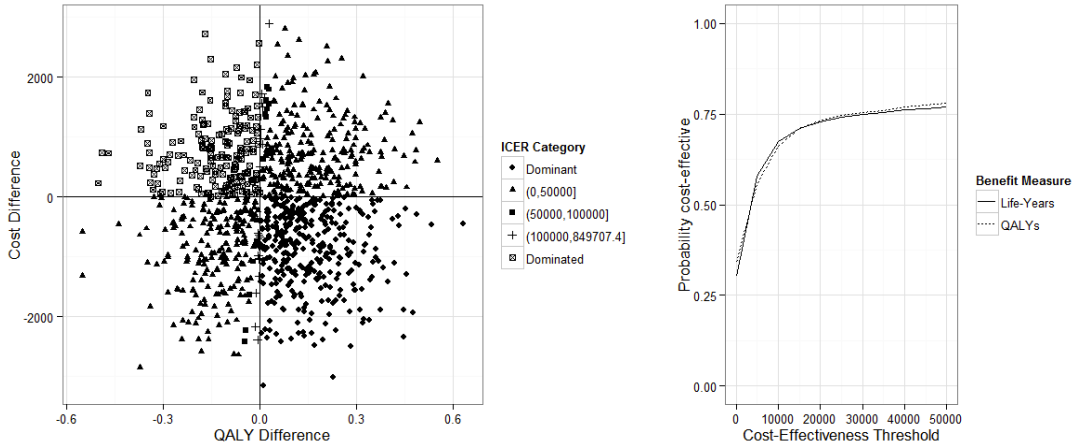
File: ICERS.png

Cost-effectiveness plane (left) and cost-effectiveness acceptability curve (right).

Case 1



Case 2



Each point on the cost-effectiveness plane represents the incremental cost (Y-axis) and incremental QALYs (X-axis) of with-testing versus without-testing *for one simulation*. Since we ran 1000 simulations in both cases, both graphs have 1000 points. As described in the Discussion of the manuscript, each simulation in the microsimulation model incorporates variation within the population. The wide range across simulations that all use the same parameter set comes from the individual-level variation in covariate values across simulations. The cost-effectiveness acceptability curves show that of the 1000 simulations, almost all Case 1 ICERs and about 75% of Case 2 ICERs fall below $50K/QALY.

Troubleshooting

After clicking "Run Analysis," the viewport to R stays blank but you see an "Analysis completed" message.

Correct Settings -> Path to R. The location for R on your machine is not correct.

Clicking "setup.exe" does not start CANTRANce.

Use the Program Manager to uninstall CANTRANce, and then re-install using "setup.exe."

Other problems

Close CANTRANce and re-start the application. If the issue remains unresolved, contact kurian@uw.edu.

CHAPTER 3: PROJECTING BENEFITS AND HARMS OF NOVEL CANCER SCREENING BIOMARKERS: A STUDY OF PCA3 AND PROSTATE CANCER

Background

Research to develop biomarkers for the early detection of cancers has expanded rapidly since the identification over twenty years ago of the now widely used cancer antigen 125 (CA-125) for ovarian cancer and prostate-specific antigen (PSA) for prostate cancer.[1] Continuing advances in molecular technology promise to widen the realm of potential biomarker candidates.[2–4] The Early Detection Research Network (EDRN) was established by the National Cancer Institute to facilitate collaboration between the increasingly numerous research groups studying new cancer biomarkers. EDRN researchers are currently studying hundreds of biomarkers spanning a wide range of cancers.[5,6]

One of the promising biomarkers under study by EDRN and others is prostate cancer antigen 3 (PCA3), a gene whose messenger RNA is overexpressed in prostate cancer tissue. Several studies have found that PCA3 increases the diagnostic accuracy of prostate cancer risk prediction models and informs the necessity for repeat biopsies.[7–14] A recent EDRN validation study concluded that using PCA3 in conjunction with PSA can improve the performance of PSA-based detection.[15] Given the accumulation of promising findings, there is growing interest in considering PCA3 as a screening tool.

As with all cancer biomarkers, advancing PCA3 from a promising biomarker to a clinically useful screening tool will require additional phases of research. Biomarker development generally progresses through several stages.[16] Initial exploratory studies are followed by the development and validation of clinical assays for the most promising biomarkers, using specimens from individuals known to have cancer and those known to be cancer-free to determine the diagnostic properties of the biomarker. Later phases of biomarker development focus on determining screening costs in the prospective setting and benefit in

terms of disease-specific mortality reduction. These phases are costly, take years to complete, and are still in the early stages for PCA3.[17]

If downstream outcomes were known earlier, these could be used to prioritize among candidate biomarkers and guide research. In this article we show how simulation modeling can link early-phase biomarker data with information on cancer natural history to project downstream population outcomes for cancer biomarkers with measured diagnostic properties. We use the case of PCA3 as a biomarker for prostate cancer. We extend an existing model of PSA growth and prostate cancer natural history[18] to incorporate PCA3 distributions based on the EDRN validation study of PCA3.[15] We then examine a range of screening strategies involving triggers for biopsy that depend on both PCA3 and PSA levels. We evaluate performance of the candidate screening strategies in terms of test results and, ultimately, in terms of overdiagnosis and lives saved. In our development, we also address general principles regarding modeling the population impact of a new biomarker for early detection of cancer.

Methods

EDRN validation study of PCA3 in prostate cancer early detection

Data on the joint distribution of PCA3, PSA, age, and prostate cancer status were obtained from a multicenter EDRN validation study of PCA3.[15] Participants were men aged 18 and over scheduled for biopsy between December 2009 and June 2011 for reasons including elevated or rising PSA, abnormal digital rectal exam (DRE), and/or low percent free PSA. PCA3 was measured in urine specimens collected following a DRE and recorded as the conventional score of the ratio of PCA3 mRNA to PSA mRNA multiplied by 1000. Cancer status and Gleason grade were determined from the prostate biopsy.

Natural history model

We used a model of PSA growth and prostate cancer developed as part of the National Cancer Institute's Cancer Intervention and Surveillance Modeling Network (CISNET) consortium.* This model links PSA growth with tumor development and progression.[18] PSA growth is slow for healthy men, faster for men following onset of a low-to-moderate grade (Gleason score 7 and below) preclinical cancer, and faster still following onset of high-grade (Gleason score 8-10) cancer. Cancer cases with faster PSA growth have shorter times to metastasis and non-screen diagnosis. The rates of PSA growth are based on data from the Prostate Cancer Prevention Trial,[19,20] which screened 9,459 men in the control group annually for up to 7 years. The natural history parameters specifying transition rates between disease states are estimated so that the model replicates prostate cancer incidence in the Surveillance, Epidemiology, and End Results (SEER) program** from 1980-2000 (Supplementary Table). The calibrated model is used to simulate life histories representative of the U.S. male population. These life histories include age and grade at onset of a biopsy-detectable preclinical tumor, grade-specific PSA trajectory, and age and stage at clinical diagnosis in the absence of screening.

Adding PCA3 to the model

To extend the simulated life histories to include individual PCA3 levels, we required information on (a) the distribution of PCA3 among prostate cancer cases and non-cases, (b) the correlation between PCA3 and PSA, (c) any association between PCA3 and disease characteristic like grade and, ideally, (d) knowledge of how PCA3 grows over a man's lifetime.

Consistent with other studies,[9,21–23] the EDRN validation study indicates that PCA3 varies by cancer status but is only weakly correlated with PSA and age.[15] We thus identified distributions for PCA3 for cancer cases and non-cases that matched the observed EDRN data but that were independent of PSA and age. Of possible lognormal and exponential distributions,

* http://cisnet.cancer.gov
** http://seer.cancer.gov

exponential distributions matching the medians in each of the cancer case and non-case groups most closely approximated the observed sensitivity and specificity of PCA3 at selected cutoffs (Table 3.1).

In the absence of longitudinal data on PCA3, we initially assumed that PCA3 discretely elevates at the time of preclinical disease onset and then stays constant. The pre- and post-onset values of PCA3 were drawn from the distributions of biopsy-negative and biopsy-positive subjects in the EDRN study, respectively.

Simulated populations

We simulated two populations with extended life histories that include PCA3 levels at all time points of interest. The first, which we call our validation population, represents a population similar to men in the EDRN study: we simulated 1 million men aged 27–86 in 1996-2000 and sampled 10,000 of those referred for biopsy according to the joint distribution of cancer status, PSA, and age at biopsy observed in the EDRN trial. The years 1996-2000 reflect the most recent years to which the natural history model is calibrated. We evaluated our integration of PCA3 into the natural history model by comparing the sensitivity and specificity of several combination PSA-PCA3 screening strategies in the validation population to the values observed in the EDRN data. We then studied population-level screening strategies in a simulated US population cohort of 10,000 men aged 50 in 2000, which we call our projected population.

Screening protocols

We simulated screening every two years in the projected population between ages 50 and 74 under eleven different screening strategies. Our nomenclature for the strategies specifies the joint cutoffs for PSA and PCA3 that define positive screening tests. The PSA-only strategy is our "base case" and uses the standard cutoff of 4.0 ng/mL for biopsy referral. We labeled this strategy *PSA(4)+PCA3(0)* since it does not include a PCA3 cutoff for biopsy referral. Based on

consultation with EDRN investigators, we evaluated two types of PSA-PCA3 combination strategies designed to improve specificity and reduce overdiagnosis. In the first type, men are referred to biopsy only if their PSA is above 4.0 ng/mL **and** their PCA3 is above a certain threshold, in an effort to "rule out" false positive tests arising from men with elevated PSA. We examined several PCA3 thresholds between 20 and 40, and these strategies are labeled *PSA(4)+PCA3(threshold)*. The second type is similar except that it additionally refers men to biopsy if their PSA is greater than 10.0 ng/mL, regardless of their PCA3. We examined the same thresholds between 20 and 40 and label these strategies *PSA(4,10)+PCA3(threshold,0).*

In all screenings, we assumed perfect compliance with biopsy referral and perfect sensitivity of the biopsy to detect existing tumors.

Prostate cancer survival in the absence of screening

The natural history model assumes that all non-metastatic diagnosed cases elect treatment according to age-, stage- and grade-specific distributions observed among SEER cases (in 2004, overall 43% chose radiation, 36% chose surgery, and 21% chose conservative management).[18] In the absence of screening, prostate cancer survival is based on age- and grade-specific survival from untreated cases reported in SEER in 1983-1986, just prior to the adoption of the PSA test for screening improved by a hazard ratio of 0.62 to for cases who elect surgery or radiation (Supplementary Table).[24] Other-cause mortality is independently generated using US life tables. When the age at prostate cancer death precedes the age at other-cause death, the death is attributed to prostate cancer.

Prostate cancer survival in the presence of screening

We modeled screening benefit with a cure model, which posits cases that would have died of cancer in the absence of screening have a probability of being cured of cancer if detected early. Cure models in the literature typically specify the probability of being cured as constant or a

function of the lead time (LT), the time by which diagnosis is advanced by screening.[25] We chose the latter approach to reflect the intuition that earlier detection may confer more survival benefit. When LT=0, screen detection occurs at the same time it would have in the absence of screening, so the probability of cure should be 0, i.e. survival should be equal to that observed in the absence of screening. As LT increases, we assume that the probability of cure increases but eventually plateaus, since there may be a limit to the marginal benefit of earlier detection. These properties are achieved by an exponential function of LT that has a single parameter $a > 0$:

$$\Pr(cure) = 1 - e^{-a(LT)}$$

We chose a value for *a* that allowed us to approximate the prostate cancer mortality results from European Randomized study of Screening for Prostate Cancer (ERSPC).[26] We screened our projected population three times, once every 4 years between ages 60–71, following the ERSPC's mean age at entry of 60.8, average inter-screen interval of 4 years, median follow-up of 11 years and PSA cutoff of 3.0 ng/mL. In targeting the trial's mortality rate ratio of 0.71 (estimated in the absence of screening noncompliance and selection bias), we found that a value of *a* = 0.2 for our cure function allowed us to approximate the ERSPC trial. The resulting cure function predicts a 33% probability of cure when LT is 2 years, a 63% probability of cure when LT is 5 years, and an 86% probability of cure when LT is 10 years. We simulated a cure status for each individual according to the resulting cure function. For individuals that were cured, the date of death was set to their date of other-cause death. Men who were not cured retained their original survival in the absence of screening.

Population outcomes

We projected both screening and mortality outcomes for the projected population. Screening outcomes include true and false positive tests, sensitivity, and specificity. In men whose disease would never manifest clinically before their death from other causes, true positive tests were

considered overdiagnoses and are reported separately. We collected mortality outcomes until age 100. Among men whose disease would manifest clinically and who consequently may be saved by screening, we report the number of prostate cancer deaths and lives saved relative to the number under no screening. Finally, we report the number of "unnecessary biopsies" as the sum of false positives and overdiagnoses. All results reported are the average over 100 simulation runs.

Sensitivity Analyses

We conducted two sensitivity analyses to examine alternative modeling assumptions. In the first, we investigated a PCA3 growth model in which PCA3 grows over time instead of staying at a constant level post-onset. In the growth model, PCA3 grows from the pre-onset level at a constant annual rate, once onset has occurred. We used the validation population to determine an annual growth rate for the PCA3 growth model that yielded a mean PCA3 among biopsy-positive cases similar to the mean PCA3 among cancer cases in the EDRN study.

In the second sensitivity analysis, we retained the constant PCA3 model but allowed PCA3 distributions to vary by grade. Based on literature suggesting a correlation between PCA3 and grade,[9,23,27,28] we assigned lower median PCA3 levels for low-grade (LG, Gleason 6) men and higher median levels for medium to high-grade (M-HG, Gleason 7-10) men under several scenarios that preserved the observed overall PCA3 median observed in the EDRN study. To adapt the natural history model accordingly, we partitioned Gleason scores 7 and below in the simulated population into Gleason score 6 versus 7 based on PSA growth rates (Supplementary Table), then combined Gleason score 7 and 8-10.

Results

EDRN validation study data

Biopsies were performed on 859 men in the EDRN validation study, of which 38% were positive. Age ranged from 27 to 86, with 46% aged 55–64 and 33% aged 65–74. PSA levels ranged between 0.20 and 309 ng/mL. Mean and median PCA3 in the study population were each over two times higher in biopsy-positive subjects than in biopsy-negative subjects (Table 3.1). As the PCA3 cutoff increased from 20 to 60, specificity increased from 56% to 89% and sensitivity decreased from 78% to 42%. PCA3 levels were not significantly different for men with low Gleason grade (6) and moderate to high grade (7-10); median PCA3 values in each group were 50 and 49 respectively.

Simulated populations

In the projected population of men aged 50 in the year 2000, thirty-six percent have preclinical onset in their lifetime (cases), beginning on average at age 65. Of cases, 38%, or 14% of the total population, would progress to clinical disease in their lifetime in the absence of screening (clinical cases). Among clinical cases, 48% have onset before age 60, preclinical disease lasts on average 14 years, and mean age at prostate cancer death in the absence of screening is 84. The remaining 62% of cases are nonclinical cases. These natural history summary measures are consistent with reported incidence for the US population and do not differ substantially from other independently developed models of prostate cancer natural history.[29]

PCA3 distributions in the simulated populations reasonably approximate those observed in the EDRN study (Table 3.1). The simulated means closely reproduce the observed means, and the mean sensitivities and specificities at three pre-selected PCA3 cutoffs fall within 3% of the observed values. In addition, the sensitivity and specificities of several combination PSA-PCA3 strategies in the validation population reasonably reproduce those observed in the EDRN study (Table 3.1).

When we used the validation population to determine a reasonable annual growth rate for the PCA3 growth model, we found that an annual growth rate of 8% yielded a mean PCA3

among biopsy-positive cases in the validation population of 68.1, similar to the trial mean among cancer cases. The median of 38.1 was lower than the observed median of 49.6, but with the constant annual growth rate constraint, no value resulted in closer fits to both the observed mean and median. When the 8% growth rate was applied to the population cohort, the mean and median PCA3 values among cancer cases were 54.4 and 32.5 at age 60 and 74.4 and 39.2 at age 70.

Screening and mortality outcomes

The PSA-PCA3 combination strategy that most reduces overdiagnosis relative to the base case PSA-based strategy is PSA(4)+PCA3(40), which cuts unnecessary biopsies by over 70% (Table 3.2). However, it also reduces the number of lives saved from 172 to 99. The PSA(4,10) strategies, which allow men with PSA above 10.0 ng/ml to test positive regardless of PCA3, all reduce overdiagnoses and unnecessary biopsies less effectively than their PSA(4) counterparts but save substantially more lives compared to the base case. For example, PSA(4,10)+PCA3(35,0) decreases overdiagnoses by 25% and unnecessary biopsies by 50% while preserving about 85% of lives saved compared to the standard PSA-based strategy, whereas PSA(4)+PCA3(35) reduces overdiagnoses by about 40% and unnecessary biopsies by 70% but only retains 60% of lives saved. Findings are very similar using the PCA3 growth model, with slightly greater overdiagnosis and slightly fewer lives saved for each strategy (data not shown).

The results are consistent with the principle that higher sensitivity generally translates to more lives saved *and* more overdiagnosed cases. Screening with PSA(4,10)+PCA3(35,0) detects 1,030 men with prostate cancer whereas screening with PSA(4)+PCA3(35) only detects 745, for example. The additional cases detected are a mixture of cases who can be saved by screening and overdiagnoses, so both lives saved and overdiagnoses increase with higher sensitivity. Conversely, because the combination PSA-PCA3 strategies were selected to

improve specificity and reduce overdiagnosis, they also reduce lives saved compared to the base case PSA-based strategy.

If PCA3 levels are not correlated with any measure of disease aggressiveness, increasing overall sensitivity amounts to increasing sensitivity similarly among both cases that would be detected in the absence of screening and overdiagnosed cases. Allowing PCA3 levels to be higher in M-HG than in LG cases (Table 3.3, select screening strategies), we find that as the correlation between PCA3 levels and grade strengthens, the sensitivity to detect M-HG cases increases while the sensitivity to detect LG cases declines. As expected, this translates into a noticeable decline in the frequency of overdiagnosis, with little change in lives saved.

Discussion

The recent recommendation against PSA screening by the US Preventive Services Task Force[30] underscores the potential for screening to induce harm as well as benefit and the need to identify screening biomarkers with favorable harm-benefit profiles as early in the development phase as possible. Early biomarker studies yield information about sensitivity and specificity of new biomarkers but not about the harm-benefit tradeoff. In this article, we use modeling to extend an EDRN validation study that found that adding PCA3 to PSA screening would reduce unnecessary prostate cancer biopsies.[15] We show how modeling may be used to project the outcomes of adding PCA3 to PSA screening using rules designed to improve test specificity while preserving sensitivity as much as possible.

We found that combination PSA-PCA3 strategies were able to substantially reduce overdiagnoses and false positive tests compared to a base case PSA-based strategy. However, approaches that most sharply reduced harms also reduced benefit. We identified several strategies that substantially reduce harms while preserving the majority of lives saved relative to the standard PSA-only strategy. Given the current emphasis on reducing the harms of prostate cancer screening, these may be valuable candidates for the future. Our results also

demonstrate that correlation with disease grade is a particularly desirable feature in screening biomarkers.

The EDRN validation study included a mixture of men presenting for initial biopsy and men returning for a repeat biopsy. Risk of cancer detection and PSA sensitivity tend to be lower in the repeat biopsy setting. We did not use this information in the model, but instead only used the PCA3 data for cancer cases and noncases. Policies could be tailored to the biopsy setting. A recent analysis using data from the same EDRN cohort investigated using PCA3 differently in these two settings, namely to rule in cancer (at PCA3>60) among men who had never previously been biopsied and to rule out cancer (at PCA3<20) among men returning for a repeat biopsy.[15] Rather than considering initial and repeat biopsies separately, our goal was to use modeling to identify a single combination rule that improved specificity and reduced overdiagnosis in both settings while preserving survival benefits. The modeling framework developed here could also be used to study the downstream consequences of rules such as the ones considered in the recent EDRN study.

Our analysis is subject to several limitations. Our results hinge upon the relationship between the natural history of prostate cancer and the joint trajectory of PSA and PCA3 over time. Following current knowledge on PCA3, we explored only two simple PCA3 trajectories that had no explicit correlation with PSA. Future studies may reveal alternative longitudinal patterns for PCA3. When we correlated PCA3 with grade, we were able to distinguish the impact of different PCA3 levels for only two grade groups, low-grade (Gleason 6) and moderate-high grade (Gleason 7-10). In addition, our method of incorporating PCA3 into prostate cancer screening represents only one possible approach. We used PCA3 in a logic rule with PSA and chose commonly cited thresholds for positive test status, but the literature on optimal cutoffs for PCA3 is not conclusive.[17] Moreover, the EDRN study and others have shown that adding PCA3 as a predictor to regression models of prostate cancer risk improves diagnostic accuracy in terms of the area under the ROC curve,[9,11,15,28,31,32] indicating that there is a continuum of risk

across levels of PCA3. We did not consider non-biomarker risk factors in modeling screening; doing so could alter the diagnostic properties and outcomes of tests based on PSA and PCA3. Finally, our model for the survival benefit of detection does not allow for early detection to extend survival rather than cure individuals of cancer. Since we chose to evaluate screening based on number of lives saved rather than survival time, this choice did not substantively impact our results, but it is an important consideration in models of cancer screening.

While our particular model reflects specific choices relevant to PCA3, the general modeling process used in this paper can serve as a prototype for modeling the impact of early detection biomarkers. We show that two sub-models are necessary: a model of how the new biomarker evolves with the natural history of disease and a model for the survival implications of early detection. The first sub-model requires data on the distribution of the new biomarker in cancer cases and non-cases, an assumption of its longitudinal course, and its correlation with existing biomarkers and disease characteristics. The second sub-model determines how early detection with a biomarker affects survival. We modeled the survival benefit using an approach that links benefit (in terms of probability of cure) to the timing of detection. If the same probability of cure was applied to all early diagnoses regardless of the timing of detection, the new biomarker would only detect additional cases.

Our work highlights the principle that simply improving sensitivity or specificity will not necessarily translate into improved population outcomes. In diseases like prostate cancer, increased sensitivity can have advantages and disadvantages, yielding more overdiagnoses, while at the same time reducing disease-specific deaths. Given that the process of advancing biomarkers from discovery to clinical approval can take many years, this type of modeling can support the process of developing and prioritizing early detection biomarkers in cancer.

References

1. Chatterjee, S. K. & Zetter, B. R. Cancer biomarkers: knowing the present and predicting the future. *Future Oncol. Lond. Engl.* **1,** 37–50 (2005).
2. Srinivas, P. R., Kramer, B. S. & Srivastava, S. Trends in biomarker research for cancer detection. *Lancet Oncol.* **2,** 698–704 (2001).
3. Manne, U., Srivastava, R.-G. & Srivastava, S. Recent advances in biomarkers for cancer diagnosis and treatment. *Drug Discov. Today* **10,** 965–976 (2005).
4. Roukos, D. H., Murray, S. & Briasoulis, E. Molecular genetic tools shape a roadmap towards a more accurate prognostic prediction and personalized management of cancer. *Cancer Biol. Ther.* **6,** 308–312 (2007).
5. Srivastava, S. & Kramer, B. S. Early detection cancer research network. *Lab. Investig. J. Tech. Methods Pathol.* **80,** 1147–1148 (2000).
6. National Cancer Institute. Objectives — EDRN Public Portal. at <http://edrn.nci.nih.gov/about-edrn/objectives/>
7. Auprich, M. *et al.* Contemporary role of prostate cancer antigen 3 in the management of prostate cancer. *Eur. Urol.* **60,** 1045–1054 (2011).
8. Tombal, B. *et al.* Biopsy and treatment decisions in the initial management of prostate cancer and the role of PCA3; a systematic analysis of expert opinion. *World J. Urol.* (2011). doi:10.1007/s00345-011-0721-0
9. De la Taille, A. *et al.* Clinical evaluation of the PCA3 assay in guiding initial biopsy decisions. *J. Urol.* **185,** 2119–2125 (2011).
10. Marks, L. S. *et al.* PCA3 molecular urine assay for prostate cancer in men undergoing repeat biopsy. *Urology* **69,** 532–535 (2007).
11. Deras, I. L. *et al.* PCA3: a molecular urine assay for predicting prostate biopsy outcome. *J. Urol.* **179,** 1587–1592 (2008).
12. Hessels, D. & Schalken, J. A. The use of PCA3 in the diagnosis of prostate cancer. *Nat. Rev. Urol.* **6,** 255–261 (2009).
13. Rigau, M. *et al.* A three-gene panel on urine increases PSA specificity in the detection of prostate cancer. *The Prostate* **71,** 1736–1745 (2011).
14. Durand, X., Moutereau, S., Xylinas, E. & de la Taille, A. Progensa™ PCA3 test for prostate cancer. *Expert Rev. Mol. Diagn.* **11,** 137–144 (2011).
15. Wei, J. T. *et al.* Can urinary PCA3 supplement PSA in the early detection of prostate cancer? *J. Clin. Oncol.* (2014).
16. Pepe, M. S. *et al.* Phases of Biomarker Development for Early Detection of Cancer. *J. Natl. Cancer Inst.* **93,** 1054 –1061 (2001).
17. Roobol, M. J. Contemporary role of prostate cancer gene 3 in the management of prostate cancer. *Curr. Opin. Urol.* **21,** 225–229 (2011).
18. Gulati, R., Inoue, L., Katcher, J., Hazelton, W. & Etzioni, R. Calibrating disease progression models using population data: a critical precursor to policy development in cancer control. *Biostatistics* **11,** 707 –719 (2010).
19. Thompson, I. M. *et al.* The influence of finasteride on the development of prostate cancer. *N. Engl. J. Med.* **349,** 215–224 (2003).
20. Etzioni, R. D. *et al.* Long-term effects of finasteride on prostate specific antigen levels: results from the prostate cancer prevention trial. *J. Urol.* **174,** 877–881 (2005).
21. Roobol, M. J. *et al.* Performance of prostate cancer antigen 3 (PCA3) and prostate-specific antigen in Prescreened men: reproducibility and detection characteristics for prostate cancer patients with high PCA3 scores (≥ 100). *Eur. Urol.* **58,** 893–899 (2010).

22. Shappell, S. B. *et al.* PCA3 urine mRNA testing for prostate carcinoma: patterns of use by community urologists and assay performance in reference laboratory setting. *Urology* **73,** 363–368 (2009).
23. Nakanishi, H. *et al.* PCA3 molecular urine assay correlates with prostate cancer tumor volume: implication in selecting candidates for active surveillance. *J. Urol.* **179,** 1804–1809; discussion 1809–1810 (2008).
24. Bill-Axelson, A. *et al.* Radical prostatectomy versus watchful waiting in early prostate cancer. *N. Engl. J. Med.* **364,** 1708–1717 (2011).
25. Wever, E. M., Draisma, G., Heijnsdijk, E. A. M. & de Koning, H. J. How does early detection by screening affect disease progression? Modeling estimated benefits in prostate cancer screening. *Med. Decis. Mak. Int. J. Soc. Med. Decis. Mak.* **31,** 550–558 (2011).
26. Schröder, F. H. *et al.* Prostate-Cancer Mortality at 11 Years of Follow-up. *N. Engl. J. Med.* **366,** 981–990 (2012).
27. Van Poppel, H. *et al.* The relationship between Prostate CAncer gene 3 (PCA3) and prostate cancer significance. *BJU Int.* (2011). doi:10.1111/j.1464-410X.2011.10377.x
28. Haese, A. *et al.* Clinical utility of the PCA3 urine assay in European men scheduled for repeat biopsy. *Eur. Urol.* **54,** 1081–1088 (2008).
29. Gulati, R. *et al.* What If I Don't Treat My PSA-Detected Prostate Cancer? Answers from Three Natural History Models. *Cancer Epidemiol. Biomarkers Prev.* **20,** 740 –750 (2011).
30. Chou, R. *et al.* Screening for Prostate Cancer: A Review of the Evidence for the U.S. Preventive Services Task Force. *Ann. Intern. Med.* (2011). doi:10.1059/0003-4819-155-11-201112060-00375
31. Ankerst, D. P. *et al.* Predicting prostate cancer risk through incorporation of prostate cancer gene 3. *J. Urol.* **180,** 1303–1308; discussion 1308 (2008).
32. Aubin, S. M. J. *et al.* Prostate cancer gene 3 score predicts prostate biopsy outcome in men receiving dutasteride for prevention of prostate cancer: results from the REDUCE trial. *Urology* **78,** 380–385 (2011).
33. Etzioni, R. *et al.* The prostate cancer conundrum revisited: treatment changes and prostate cancer mortality declines. *Cancer* **118,** 5955–5963 (2012).
34. Cooperberg, M. R., Vickers, A. J., Broering, J. M. & Carroll, P. R. Comparative risk-adjusted mortality outcomes after primary surgery, radiotherapy, or androgen-deprivation therapy for localized prostate cancer. *Cancer* **116,** 5226–5234 (2010).
35. Schröder, F. H. *et al.* Screening and Prostate-Cancer Mortality in a Randomized European Study. *N. Engl. J. Med.* **360,** 1320–1328 (2009).
36. National Center for Health Statistics. *Vital Statistics of the United States, Volume II: Mortality, Part A*. (Government Printing Office).

Tables

Table 3.1

Mean, median, and variance of PCA3, and sensitivity and specificity of various PCA3 and PSA cutoffs in the EDRN data and the simulated populations (validation and projected). The simulated values for PCA3 are derived from the exponential distributions matching the observed PCA3 medians.

		EDRN Data	**Simulated Populations**
PCA3 Mean	Noncancer	29.3	25.6
	Cancer	68.7	71.8
PCA3 Median	Noncancer	17.8	17.8
	Cancer	49.8	49.8
PCA3 Standard	Noncancer	1869.8	655.8
Deviation	Cancer	4487.0	5161.9
Sensitivity	PCA3>20	0.78	0.76
	PCA3>35	0.62	0.61
	PCA3>60	0.42	0.43
	PSA(4)+PCA3(0)*	0.79	0.81
	PSA(4)+PCA3(20)*	0.63	0.60
	PSA(4,10)+PCA3(20,0)*	0.66	0.65
Specificity	PCA3>20	0.56	0.54
	PCA3>35	0.77	0.75
	PCA3>60	0.89	0.90
	PSA(4)+PCA3(0)*	0.28	0.29
	PSA(4)+PCA3(20)*	0.68	0.67
	PSA(4,10)+PCA3(20,0)*	0.64	0.59

*Numbers in "Simulated Population" column refer to validation population only. See methods section for key to nomenclature.

Table 3.2

Screening and prostate cancer (PC) mortality outcomes in the projected population (N=10,000) under different screening strategies. Lives saved are relative to no screening. Unnecessary biopsies are the sum of false positive tests and overdiagnoses.

	Strategy	True Positive Tests	False Positive Tests	Overdiagnoses	Unnecessary Biopsies	PC Deaths	Lives Saved
1	PSA(4)+PCA3(0)	1217	6912	385	7297	180	173
2	PSA(4)+PCA3(20)	920	3159	291	3450	222	131
3	PSA(4)+PCA3(25)	858	2599	272	2871	231	122
4	PSA(4)+PCA3(30)	801	2132	254	2386	240	113
5	PSA(4)+PCA3(35)	747	1751	237	1988	247	106
6	PSA(4)+PCA3(40)	697	1441	221	1662	254	99
7	PSA(4,10)+PCA3(20,0)	1101	4220	327	4547	194	159
8	PSA(4,10)+PCA3(25,0)	1076	3818	315	4133	198	155
9	PSA(4,10)+PCA3(30,0)	1054	3481	304	3785	200	153
10	PSA(4,10)+PCA3(35,0)	1032	3206	294	3500	203	150
11	PSA(4,10)+PCA3(40,0)	1013	2982	284	3266	206	146
12	No Screening					353	

Table 3.3

Screening and prostate cancer (PC) mortality outcomes in the projected population (N=10,000) when PCA3 varies by grade, for select strategies from Table 2. Grade-specific sensitivities refer to the PCA3 test properties alone. Grade is either low grade (LG) or moderate to high-grade (M-HG).

Strategy	LG Median PCA3	M-HG Median PCA3	LG Sensitivity of PCA3	M-HG Sensitivity of PCA3	LG Cases Detected	M-HG Cases Detected	Over-diagnoses	PC Deaths	Lives Saved
PSA(4)+PCA3(20)	50	50	0.758	0.758	531	389	291	223	130
	30	85	0.630	0.850	443	437	272	223	130
	20	103	0.500	0.874	353	449	241	228	125
PSA(4)+PCA3(35)	50	50	0.616	0.616	431	316	237	247	105
	30	85	0.445	0.752	312	386	211	246	107
	20	103	0.297	0.790	211	406	178	250	103

Supplementary Table

Data and assumptions used to construct each element of the natural history model of prostate cancer and the extension incorporating PCA3.

Component	Element	Depends on	Dataset	Publication(s)
Natural history model	Pre-onset PSA growth	age	PCPT	Gulati et al, 2010[18]
	Post-onset PSA growth	age, age at onset, grade	PCPT	Gulati et al, 2010[18], Etzioni et al, 2012[33]
	Cancer onset	age	SEER 9	Gulati et al, 2010[18]
	Cancer metastasis	PSA	SEER 9	Gulati et al, 2010[18]
	Clinical detection	PSA	SEER 9	Gulati et al, 2010[18]
	Treatment patterns	age, year, stage, grade	SEER 9	Etzioni et al, 2012[33]
	Prostate cancer survival for unscreened untreated cases	age, stage, grade	SEER 9	Etzioni et al, 2012[33]
	Prostate cancer survival benefit for curative treatment	(local-regional stage only)	SPCG-4 trial, CaPSURE, ERSPC	Bill-Axelson et al., 2011[24], Cooperberg et al, 2010[34], Schröder et al, 2012[35]
	Overall survival	birth year	Vital Statistics of the US	National Center for Health Statistics, various years[36]
PCA3 extension	PCA3 levels	prostate cancer status	EDRN trial	Wei et al, 2014[15]
	PCA3 trajectory	prostate cancer onset	Assumption[1]	
Screen-detected survival	Prostate cancer survival benefit of screening (probability of cure)	lead time	Calibrated assumption[2]	

[1]Main analysis assumes that PCA3 discretely elevates at prostate cancer onset. Sensitivity analysis allows PCA3 to grow at an annual rate of 8%.

[2]We assume that individuals are cured with probability $1 - e^{-0.2(LT)}$, where LT is the time by which screen detection advanced diagnosis from the time of clinical diagnosis. See Methods for details.

CHAPTER 4: THE IMPACT OF TREATMENT ADVANCES ON THE MORTALITY RESULTS OF CANCER SCREENING TRIALS

Background

Early detection of cancer used to be viewed as an unequivocal advantage, but academic and public dialogue are now polarized as to whether the benefits of population screening programs outweigh their harms.[1–5] Central to the debate is the evidence provided by randomized controlled trials (RCTs) on the reduction in cancer mortality attributable to screening. Policy panels are mandated to prioritize RCTs when developing guidelines, as indicated by the US Preventive Services Task Force's assignment of RCTs as "Level I evidence" in their hierarchy of research designs.[6] When results from RCTs conflict, as in breast and prostate cancer screening trials,[7,8] forming guidelines becomes controversial.

Recent attempts to reconcile the diverging conclusions of screening trials have unearthed a particularly intractable issue regarding the relevance of treatments administered in cancer screening RCTs. [9–12] These RCTs span decades, and during that time, technological advances may render the trial treatments out of date. The benefit of early detection of cancer is inextricably linked to the effectiveness of treatment in curing or delaying mortality from the disease.[13] Treatment thus plays an important role in producing trial results.

Recognizing this link between treatment and screening, some breast cancer researchers hypothesize that contemporary treatments have decreased the mortality impact of screening mammography.[14–16] Many of the mammography trials, initiated in the 1960s and 1970s,[17] predated the advent of adjuvant multiagent chemotherapy and tamoxifen that disseminated into the population in the 1980s and 1990s and are now standards of care.[18] The availability of these therapies is cited as a potential reason why the most recently conducted trial, the Canadian National Breast Screening Study, did not find a significant effect of screening[3,8,19] unlike the 15-20% mortality reduction estimated from meta-analyses of all trials.[14,17,20] The recent Cochrane

Review and Swiss Medical Board guidelines regarding mammography similarly cite advances in treatment as support for the hypothesis that the true mortality reduction due to mammography is less than that suggested by meta-analyses.[14,16]

A recent *Annals of Internal Medicine* article proposed that "The only way to know [the effect of mammography screening] for certain is to initiate a new trial in the era of contemporary screening technologies and breast cancer therapies."[15] Unfortunately, the 20 years of follow-up needed to reliably observe mortality reductions attributable to screening[21] present a problem of timing that no trial can circumvent. Treatments administered in a new trial may once again prove outdated in 20 years.[22] Moreover, it is not possible to predict the direction in which treatment advances will impact screening. When treatments improve for tumors detected early in the process of disease progression, the benefit of screening will increase. Conversely, better treatments for more advanced disease will decrease the value of screening.

When empirical studies fall short of providing the evidence needed for policy development, modeling can help. Models have been used to decompose the joint contributions of screening and treatment to trends in breast cancer mortality for the population setting,[9,23] but the particular impact of treatment advances on screening trial results has not been investigated. In this paper, we present a simple modeling framework to quantify how screening trial results change when treatments become more advanced. We apply this framework to the case of breast cancer to evaluate the hypothesis that advances in breast cancer treatment have decreased the impact of screening on breast cancer mortality.

Methods

Model Overview

The goal of screening is to detect tumors that are early enough in cancer progression to be amenable to treatment and possibly cure. Studies indicate that one of the effects of screening is

to diagnose cancers in an earlier stage[24–28] in which cancer cells are less likely to have spread compared with advanced-stage disease.

We use this “stage-shift” mechanism[29,30] to model the impact of screening on cancer survival. In the absence of screening, cancer diagnoses are termed “clinical” diagnoses, and stage at diagnosis is a mixture of “early" stage and “advanced” stage. Early-stage cases survive longer after clinical diagnosis than advanced cases. Screening makes it possible for more cases to be diagnosed in early stage, so in the screened setting, the model shifts some cases who were advanced stage at clinical diagnosis to be early stage.

The stage-shift model thus has two avenues for screening to improve survival: first, through the assumption that some diagnoses will shift from having advanced-stage survival to the longer early-stage survival, and second, through any potential advantage of early-stage treatments. However, availability of highly effective advanced-stage treatments will decrease the survival-time gap between the stages and decrease the stage-shift benefit.

Our modeling framework consists of a series of virtual screening trials that all have the same screening-induced stage shift but reflect different treatment patterns. Differences in trial outcomes will thus reflect the impact of different treatments. To apply this framework to breast cancer, we model three trials: a “Historical” trial with treatments available in the US population in the late 1970s, a “1999” trial with US population treatment frequency from 1999, and a “Perfect” trial that reflects optimal utilization of contemporary treatments. None of these trials are intended to mimic any of the mammography trials that have been conducted. Instead, the trials each examine mammography in different eras of treatment in the US population. Together, their results show the impact of treatment advances on a virtual mammography trial that would have been conducted in the US given treatment distributions corresponding to each of these time points, but under a specified stage shift that is based on the screening trials actually conducted.

Model Structure

We use microsimulation modeling to project results of the virtual screening trials. Each breast cancer trial consists of 100,000 women aged 50 in 2000. We simulate simple "life histories" in the absence of screening by projecting the dates and characteristics of the following events (see Table 4.A1 for more details):

1. *Age at clinical diagnosis of cancer* is simulated from age-specific rates of clinical diagnosis. We used breast cancer diagnosis rates from the Surveillance, Epidemiology and End Results (SEER) database[31] for the years 1975-1979, a "historic" period that predates widespread screening and contemporary treatments. We note that while screening can lead to overdiagnosis, in this model, we focus only on clinical cancer cases who would have been diagnosed in their lifetimes in the absence of screening.
2. *Stage at clinical diagnosis in the absence of screening* is assigned as early or advanced. We used historic SEER breast cancer stage distributions for 1975-1979 to inform stage distributions without screening, classifying "local" stage as early stage and "regional" and "distant" stages as advanced stage.
3. *Other stage-specific disease characteristics,* if applicable. We included breast cancer tumor receptor category, because different treatments are utilized depending on estrogen receptor (ER) and human epidermal growth factor receptor 2 (HER2) receptor status.[32] Separately for early and advanced stage, we used SEER breast cancer cases diagnosed in 2010 with known receptor status to simulate an ER positive (+) or negative/borderline (-) and HER2 positive (+) or negative/borderline (-) status.
4. *Age at cancer death, baseline*. We use exponential survival curves, one for each stage, to simulate a "baseline" time from clinical diagnosis to cancer death in the absence of death from other causes. Baseline survival reflects survival during the time period when contemporary treatments were not used. To define these survival curves for breast cancer, we averaged the mortality rates implied by the 5- and 10-year survival in each

stage group in historic SEER data (1975-1979 diagnoses), and used frequency-weighting to combine the "regional" and "distant" rates into an advanced-stage rate.

5. *Age at cancer death, treated.* Contemporary treatments are incorporated by applying hazard ratios to baseline survival to modify the time from clinical diagnosis to cancer death. The hazard ratios depend on the treatments received and their efficacies:
    a. *Distribution of treatments.* We patterned treatments in the breast cancer trials to reflect the evolution of treatments used in the US population as reported in a study of adjuvant treatment dissemination[33] (Figure 4.1). NIH recommendations evolved from treatment with multiagent chemotherapy to the addition of tamoxifen for ER+ tumors over the course of the 1970s to 2000, and clinical practices followed.[33] Accordingly, in the Historical trial, we used treatment patterns from 1975 when only chemotherapy was made available to just a fraction of advanced cases. The 1999 trial reflects the shift towards using ER status to target adjuvant tamoxifen as well as greater use of chemotherapy. The Perfect trial approximates an ideal contemporary scenario by assuming that combination chemotherapy and tamoxifen is used for ER+ cases and chemotherapy alone is usedfor ER- cases. In addition, HER2+ cases of either ER status receive trastuzumab in the Perfect trial.
    b. *Efficacies of treatments.* Treatment efficacies were derived from published meta-analyses from the Early Breast Cancer Trialists' Group (EBCTG) and the Cochrane Reviews. The most recent EBCTG reviews report disease-specific hazard ratios (HRs) for chemotherapy (average across regimens) and tamoxifen (ER+ only) as 0.775 and 0.70, respectively, for all operable breast cancers.[34,35] Cochrane reviews of trastuzumab report HRs for overall survival in HER2+ early and metastatic breast cancers as 0.66 and 0.82, respectively.[36,37] In both sets of reviews, the HRs are meta-analysis estimates of the overall impact of the

presence of the treatment compared to its absence across a range of contexts, e.g. the tamoxifen trials varied in the presence and administration of background chemotherapy. We applied the chemotherapy and tamoxifen HRs to both our early- and advanced- stage cases. For trastuzumab, we treated the HRs for overall survival as proxies for disease-specific HRs. We applied the differing early and metastatic trastuzumab HRs to our stage groupings by using the early HR for early stage and the metastatic HR for advanced stage. Finally, we assumed multiplicative effects on the HR scale for combination therapies.

6. *Age at other-cause death*. We simulate an age at other-cause death from US life tables, so that the overall mortality in our population represents that of the US.
7. *Cause of death and age at death*. The cause of death and age at death are taken from the event that occurs first, cancer or other-cause death.

In the presence of screening we subject the same individuals to an alternate scenario in which a specified proportion of advanced stage cases are shifted to early stage due to screening. The proportion shifted reflects the combined impact of screening frequency, participation, and test sensitivity. We used a stage-shift proportion of 15% for the breast cancer trials, which reflects the median of the range of decreases in the cumulative incidence of advanced-stage disease reported across 8 mammography trials.[28]

Stage-shifted cases have their age at cancer death re-generated using steps 4 & 5 with their new early-stage status. Their survival times under screening are explicitly correlated with their survival time in the absence of screening. Within each trial, both the control and screening arms receive the same stage-specific distributions of treatment.

Outcomes

Following typical screening trial protocol, we projected the cumulative incidence of breast cancer mortality in each trial's screening and control arm. We assessed the impact of screening

using the mortality rate ratio (MRR), or the ratio of mortality in the screening arm to mortality in the control arm, and the absolute risk reduction (ARR), the absolute amount by which screening reduced cumulative incidence of breast cancer mortality. We report these statistics at the 13-year follow-up point.

While the standard MRR and ARR reflect the effect of screening in a trial where participants in both arms receive the same stage- and receptor-specific treatments, our simulated trials also allow us to compare the control and screening arms across trials. We compared the control arms of the two later trials to that of the Historical trial to estimate the effect of treatment advances without screening, and we similarly compared screening arms to estimate the effect of treatment advances with screening.

Sensitivity Analyses

Under the stage-shift model, screening will be somewhat beneficial due to the stage shift alone. Additional benefit may arise when treatments are available to further extend early-stage survival. Conversely, the stage-shift benefit can be minimized if advanced-stage treatments are efficacious enough to extend advanced-stage survival to be comparable with early-stage survival. In contrast to our breast cancer example in which most treatments appear to be equally efficacious for both stages, in our first sensitivity analysis we simulated several screening trials in which we applied one treatment to each stage and varied the HRs of each stage-specific treatment from 0.25 (very beneficial) to 1 (no benefit). This allowed us to separate treatment efficacy by stage.

In a second sensitivity analysis, we explored sensitivity to the stage shift proportion by increasing the stage shift to 25%. This greater stage shift could represent increased screen detection due to better technology or more frequent screening. Finally, in a third sensitivity analysis, we examined the impact of varying the relationship between the baseline early- and

advanced-stage mortality by halving the early-stage mortality rate. This scenario could represent a major advance in local therapy alone.

Results

The projected cumulative incidence of breast cancer mortality in the trials follows the expected pattern of lower mortality in screening than control arms and lower mortality as treatments improve across trials (Figure 4.2). Cumulative mortality ranges from about 2.5% in the control arm of the Historical trial by 25 years of follow-up down to less than 1.5% in the screening arm of the Perfect trial, suggesting an absolute long-term effect of screening plus improved treatment of approximately 1% or a relative long-term effect of 40% reduction in disease-specific mortality.

The relative trajectories of the screening vs control arms in Figure 4.2 are similar across trials, and the MRRs at 13 years are almost identical after accounting for simulation uncertainty (Table 4.1). The virtual trials project an 8-10% relative decrease in mortality due to screening and a 0.05-0.07% absolute reduction. Thus, the results indicate that improved treatment as modeled across the three trial settings does not substantively affect mortality benefit associated with screening. This finding may seem counterintuitive but is consistent with the fact that the treatments modeled were systemic and comparably advantageous for both stages. An exception is trastuzumab, available in the Perfect trial, but this treatment was only relevant to the small HER2+ subpopulation.

A substantial impact of treatment advances is seen only when screening is held constant and comparisons are made across trials (Table 4.1). Regardless of whether the comparison is of control or screening arms, the across-trial relative mortality reduction due to treatment advances holding screening constant is projected as at least twice that of the within-trial reduction due to screening holding treatment constant.

Our first sensitivity analysis shows that when treatment advantages the stages differentially, there is a greater range of the within-trial MRRs. Across 16 trials with varying efficacies of early- and advanced-stage treatment, MRRs span a range of 10%, with screening reducing mortality by 3-13% (Figure 4.3). As expected, the weakest MRRs occur when the advanced treatment is highly effective but the early treatment is not, and the strongest MRRs occur in the opposite case.

When the stage shift increases to 25% as in our second sensitivity analysis, the projected mortality reductions in each trial increase but still remain similar across trials. Within-trial relative mortality reductions due to screening increase to 15-16%, and absolute reductions increase to 0.08-0.1% (Table 4.A2). In the first sensitivity analysis, more efficacious early-stage treatments have a more striking impact under the greater stage shift, with reductions up to 22% (Figure 4.A1).

Our third sensitivity analysis also projects increased mortality reductions that remain similar across trials. Increasing the gap between baseline early- and advanced-stage mortality increases breast cancer mortality reductions due to screening from 8-10% to 10-11% across trials (Table 4.A3). Mortality reductions also remain similar across the different treatment settings explored in the first sensitivity analysis; with early-stage mortality very low, even a treatment HR of 0.25 has a very minor impact on the survival curve, which minimizes the impact of varying early-stage treatment HRs (Figure 4.A2).

Discussion

Policy-makers are continuously faced with the dilemma that treatments have changed since screening RCTs were initiated. Modeling provides a way to update historical trial results by decomposing the benefit of screening into a stage shift due to screening and the added benefit of treatment. We use this approach to understand how treatment advances may impact the results of screening trials. Counter to prevailing speculation, we find that the survival benefits

conferred by improved treatment do not substantively change the projected mortality impact of mammography screening. This is because breast cancer treatments are similarly effective in both stages. The dissemination of effective treatment has also been fairly comparable for both stages.[33] As a result, we projected similar MRRs for all three trials despite major differences in population treatment distributions. These similarities persist regardless of the magnitude of the stage shift due to screening or the difference in baseline survival rates by stage.

Trial results change when treatment improvements affect different stages differentially. Our first sensitivity analysis shows how MRR differences between trials emerge as the stage-specific efficacies of treatments diverge. Better early-stage treatments increase the MRR. This effect increases with larger stage shifts, because a larger stage shift allows good early-stage treatments to impact the population more. However, the low baseline mortality rate for early-stage breast cancer sets a limit for how much early-stage treatments can improve the MRR. After a certain point, only a greater stage shift can meaningfully increase the benefit of mammography screening.

Because our framework and examples are highly simplified, it is important to examine the model's limitations in light of their potential impact on our main conclusions. We have attempted to provide projections that shed light on hypotheses regarding the impact of treatment advances on screening trial results, using a simple model that is useful despite its highly stylized representation of reality. The most influential simplification is the assumption that the stage-shift mechanism captures the essence of the benefit of early detection. Evidence for the screening-related reduction of advanced-stage cancer incidence is weaker in population studies than in screening trials, and some researchers question whether it is truly possible to detect advanced-stage tumors early.[38,39] Even if screening does reduce advanced-stage incidence, we model the survival benefit of the stage shift in a way that has precedence in the literature but cannot be confirmed biologically.

In addition, recent advances in breast cancer research have identified features of tumor biology that are highly significant to disease prognosis and therapeutic response, perhaps even more than disease stage.[40] Genomic tests have emerged that can effectively classify disease prognosis.[41] "Triple negative" breast cancer patients, whose tumors are negative for ER, HER2 and progesterone receptor, appear to have particularly poor survival and optimal treatment is unknown.[42–44] Certain tumor types may be more easily detected with existing screening technologies[45] but our model does not allow this. Further, the lack of available data on survival in screened versus clinically-detected cohorts of varying tumor characteristics prohibited us from modeling differential survival (independent differential treatment response) for different tumor types.

Our focus on modeling breast cancer in the population (and particularly our use of population baseline survival curves by stage) means that our projected mortality reductions associated with screening do not quite match those observed in the trials. Mortality benefit due screening is generally cited as approximately 20% reduction is disease-specific mortality,[28] but the model produced about a 10% reduction in mortality, given specified baseline survival rates. (We do note that other, more complex models have similarly estimated a 10% reduction in breast cancer mortality due to screening in the population setting.[9]) We also did not calibrate age at randomization or any of the life history parameters to a particular trial, but directly used historic SEER data. We only included those local therapies represented in the SEER survival curves and did not model any contemporary local therapies. In order to use published HRs of systemic treatments, we assumed that historic SEER diagnoses are comparable to the treatment trial populations in terms of treatment response. In our Perfect trial, we assumed that combination tamoxifen and chemo is optimal in ER+ women, even though recent research has identified tumor subgroups for whom chemotherapy does not improve survival.[47] We approximated survival from clinical diagnosis with an exponential curve which is likely a simplification. Finally, we did not model any "beyond stage-shift" effect of screening that may

arise from the fact that screen-detectable tumors may grow more slowly than clinically-detected early-stage tumors.[46] All these simplifications should not impact our basic conclusions, because they either affect all our virtual trials or affect only a small subgroup of cancers.

While all models are limited representations of reality, we attempt to provide a simple yet useful first framework for examining the impact of treatment advances on screening trial results. By decomposing screening into its effect on diagnoses and subsequent treatment, we provide some insight into this question for mammography screening. Our work also highlights the characteristics of treatment advances that would most likely impact screening trial results now and in the future. Our findings may help set realistic expectations regarding what questions can be answered by initiating new screening trials.

References

1. Kolata, G. Cancer Screening May Be More Popular Than Useful. *The New York Times* (2011). at <http://www.nytimes.com/2011/10/30/health/cancer-screening-may-be-more-popular-than-useful.html?_r=1&scp=6&sq=prostate%20cancer%20screening&st=cse>
2. Friedrich MJ. Debate continues on use of PSA testing for early detection of prostate cancer. *JAMA* **305,** 2273–2276 (2011).
3. Harris, R. Screening Is Only Part of the Answer to Breast Cancer. *Ann. Intern. Med.* **160,** 861–863 (2014).
4. Woolf, S. H. & Harris, R. The harms of screening: new attention to an old concern. *JAMA* **307,** 565–566 (2012).
5. Garnick, M. B. The Great Prostate Cancer Debate. *Sci. Am.* **306,** 38–43 (2012).
6. Agency for Healthcare Research and Quality. *US Preventive Services Task Force Procedural Manual*. 36 (2008). at <http://www.uspreventiveservicestaskforce.org/Home/GetFile/6/7/procmanual/pdf>
7. Moyer, V. A. & U.S. Preventive Services Task Force. Screening for prostate cancer: U.S. Preventive Services Task Force recommendation statement. Ann. Intern. Med. 157, 120–134 (2012).
8. Miller, A. B. *et al.* Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ* **348,** g366 (2014).
9. Berry, D. A. Breast cancer screening: controversy of impact. *Breast Edinb. Scotl.* **22 Suppl 2,** S73–76 (2013).
10. Etzioni, R. *et al.* Limitations of basing screening policies on screening trials: The US Preventive Services Task Force and Prostate Cancer Screening. *Med. Care* **51,** 295–300 (2013).
11. Marmot, M. G. Sorting through the arguments on breast screening. *JAMA J. Am. Med. Assoc.* **309,** 2553–2554 (2013).
12. Gøtzsche, P. C., Jørgensen, K. J., Zahl, P.-H. & Mæhlen, J. Why mammography screening has not lived up to expectations from the randomised trials. *Cancer Causes Control CCC* **23,** 15–21 (2012).
13. Feuer, E. J. Chapter 1: Modeling the Impact of Adjuvant Therapy and Screening Mammography on U.S. Breast Cancer Mortality Between 1975 and 2000: Introduction to the Problem. *JNCI Monogr.* **2006,** 2 –6 (2006).
14. Gøtzsche, P. C. & Jørgensen, K. J. Screening for breast cancer with mammography. *Cochrane Database Syst. Rev.* **6,** CD001877 (2013).
15. Jüni, P. & Zwahlen, M. It Is Time to Initiate Another Breast Cancer Screening Trial. *Ann. Intern. Med.* **160,** 864–866 (2014).
16. Biller-Andorno, N. & Jüni, P. Abolishing mammography screening programs? A view from the Swiss Medical Board. *N. Engl. J. Med.* **370,** 1965–1967 (2014).
17. Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. *Lancet* **380,** 1778–1786 (2012).
18. Sledge, G. W. *et al.* Past, present, and future challenges in breast cancer treatment. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **32,** 1979–1986 (2014).
19. Kalager, M., Adami, H.-O. & Bretthauer, M. Too much mammography. *BMJ* **348,** g1403 (2014).
20. US Preventive Services Task Force. Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. *Ann. Intern. Med.* **151,** 716–726, W–236 (2009).

21. Smith, R. A. The value of modern mammography screening in the control of breast cancer: understanding the underpinnings of the current debates. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **23,** 1139–1146 (2014).
22. Elmore, J. G. & Harris, R. P. The harms and benefits of modern screening mammography. *BMJ* **348,** g3824 (2014).
23. Berry, D. A. *et al.* Effect of screening and adjuvant therapy on mortality from breast cancer. *N. Engl. J. Med.* **353,** 1784–1792 (2005).
24. Horeweg, N. *et al.* Detection of lung cancer through low-dose CT screening (NELSON): a prespecified analysis of screening test performance and interval cancers. *Lancet Oncol.* (2014). doi:10.1016/S1470-2045(14)70387-0
25. Hankey, B. F. *et al.* Cancer surveillance series: interpreting trends in prostate cancer--part I: Evidence of the effects of screening in recent prostate cancer incidence, mortality, and survival rates. *J. Natl. Cancer Inst.* **91,** 1017–1024 (1999).
26. Pashayan, N. *et al.* Mean sojourn time, overdiagnosis, and reduction in advanced stage prostate cancer due to screening with PSA: implications of sojourn time on screening. *Br. J. Cancer* **100,** 1198–1204 (2009).
27. Fracheboud, J. *et al.* Decreased rates of advanced breast cancer due to mammography screening in The Netherlands. *Br. J. Cancer* **91,** 861–867 (2004).
28. Autier, P., Héry, C., Haukka, J., Boniol, M. & Byrnes, G. Advanced breast cancer and breast cancer mortality in randomized controlled trials on mammography screening. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **27,** 5919–5923 (2009).
29. Connor, R. J., Chu, K. C. & Smart, C. R. Stage-shift cancer screening model. *J. Clin. Epidemiol.* **42,** 1083–1095 (1989).
30. Wever, E. M., Draisma, G., Heijnsdijk, E. A. M. & de Koning, H. J. How does early detection by screening affect disease progression? Modeling estimated benefits in prostate cancer screening. *Med. Decis. Mak. Int. J. Soc. Med. Decis. Mak.* **31,** 550–558 (2011).
31. National Cancer Institute. Surveillance, Epidemiology and End Results Program. at <www.seer.cancer.gov>
32. Ludwig, J. A. & Weinstein, J. N. Biomarkers in Cancer Staging, Prognosis and Treatment Selection. *Nat. Rev. Cancer* **5,** 845–856 (2005).
33. Mariotto, A. B., Feuer, E. J., Harlan, L. C. & 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.* 7–15 (2006). doi:10.1093/jncimonographs/lgj003
34. Early Breast Cancer Trialists' Collaborative Group (EBCTCG) *et al.* Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. *Lancet* **378,** 771–784 (2011).
35. Early Breast Cancer Trialists' Collaborative Group (EBCTCG) *et al.* Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. *Lancet* **379,** 432–444 (2012).
36. Balduzzi, S. *et al.* Trastuzumab-containing regimens for metastatic breast cancer. *Cochrane Database Syst. Rev.* **6,** CD006242 (2014).
37. Moja, L. *et al.* Trastuzumab containing regimens for early breast cancer. *Cochrane Database Syst. Rev.* **4,** CD006243 (2012).
38. Autier, P. *et al.* Advanced breast cancer incidence following population-based mammographic screening. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. ESMO* **22,** 1726–1735 (2011).
39. Nederend, J. *et al.* Trends in incidence and detection of advanced breast cancer at biennial screening mammography in The Netherlands: a population based study. *Breast Cancer Res. BCR* **14,** R10 (2012).

40. Kalia, M. Personalized oncology: Recent advances and future challenges. *Metabolism* **62, Supplement 1,** S11–S14 (2013).
41. Chung, C. & Christianson, M. Predictive and prognostic biomarkers with therapeutic targets in breast, colorectal, and non-small cell lung cancers: A systemic review of current development, evidence, and recommendation. *J. Oncol. Pharm. Pract. Off. Publ. Int. Soc. Oncol. Pharm. Pract.* **20,** 11–28 (2014).
42. Ismail-Khan, R. & Bui, M. M. A review of triple-negative breast cancer. *Cancer Control J. Moffitt Cancer Cent.* **17,** 173–176 (2010).
43. Bouchalova, K. *et al.* Triple Negative Breast Cancer - BCL2 in Prognosis and Prediction. Review. *Curr. Drug Targets* **15,** 1166–1175 (2014).
44. Liedtke, C. *et al.* Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **26,** 1275–1281 (2008).
45. Chuang, S.-L. *et al.* Using tumor phenotype, histological tumor distribution, and mammographic appearance to explain the survival differences between screen-detected and clinically detected breast cancers. *APMIS Acta Pathol. Microbiol. Immunol. Scand.* **122,** 699–707 (2014).
46. Shen, Y. *et al.* Role of detection method in predicting breast cancer survival: analysis of randomized screening trials. *J. Natl. Cancer Inst.* **97,** 1195–1203 (2005).
47. Albain, K. S., Paik, S. & van't Veer, L. Prediction of adjuvant chemotherapy benefit in endocrine responsive, early breast cancer using multigene assays. *Breast Edinb. Scotl.* **18 Suppl 3,** S141–145 (2009).
48. Stockler, M., Wilcken, N. R. C., Ghersi, D. & Simes, R. J. Systematic reviews of chemotherapy and endocrine therapy in metastatic breast cancer. *Cancer Treat. Rev.* **26,** 151–168 (2000).

Tables and Figures

Table 4.1

Impact of varying treatment on the cumulative incidence of breast cancer mortality observed in a screening trial.
Results reflect 13 years of follow-up in a female cohort of size 100,000 aged 50 in 2000, with 95% uncertainty intervals across 100 simulations. In all trials, screening reduces advanced disease by 15%.

	TREATMENT GIVEN IN THE SCREENING TRIAL		
	Historical Treatment	**1999 Treatment**	**Perfect Treatment**
Cumulative incidence, control group			
Control arm	0.0075 (0.0073,0.008)	0.0059 (0.0057,0.0064)	0.005 (0.0048,0.0054)
Screening arm	0.0068 (0.0066,0.0074)	0.0054 (0.0052,0.0058)	0.0045 (0.0044,0.0049)
Effect of screening, same treatment			
MRR	0.9067 (0.9062,0.934)	0.9153 (0.9038,0.9318)	0.9 (0.8946,0.9261)
ARR	0.0007 (6e-04,8e-04)	0.0005 (5e-04,6e-04)	0.0005 (4e-04,6e-04)
Effect of treatment, without screening*			
MRR		0.7867 (0.7797,0.8091)	0.6667 (0.6512,0.6962)
ARR		0.0016 (0.0015,0.0019)	0.0025 (0.0024,0.0028)
Effect of treatment, with screening^			
MRR		0.7941 (0.7764,0.8149)	0.6618 (0.6428,0.6931)
ARR		0.0014 (0.0014,0.0017)	0.0023 (0.0022,0.0026)

* Compares control arm of the trial to control arm of the Historical Trial

^ Compares screening arm of the trial to screening arm of the Historical Trial

Figure 4.1

Treatment distributions in each of the three trials, conditional on stage at diagnosis and ER receptor status. The Historical and Contemporary 1999 distributions are derived from the Mariotto 2006 analysis of treatment dissemination in the US population in 1975 and 1999, respectively, applying their Stage I results to our early stage and their Stage II+/IIIA results to our advanced stage. The Perfect trial approximates an “optimal” scenario in which all ER+ cases receive tamoxifen and chemotherapy and all ER- cases receive chemotherapy. In addition, in the Perfect trial, all HER2+ cases additionally receive trastuzumab.

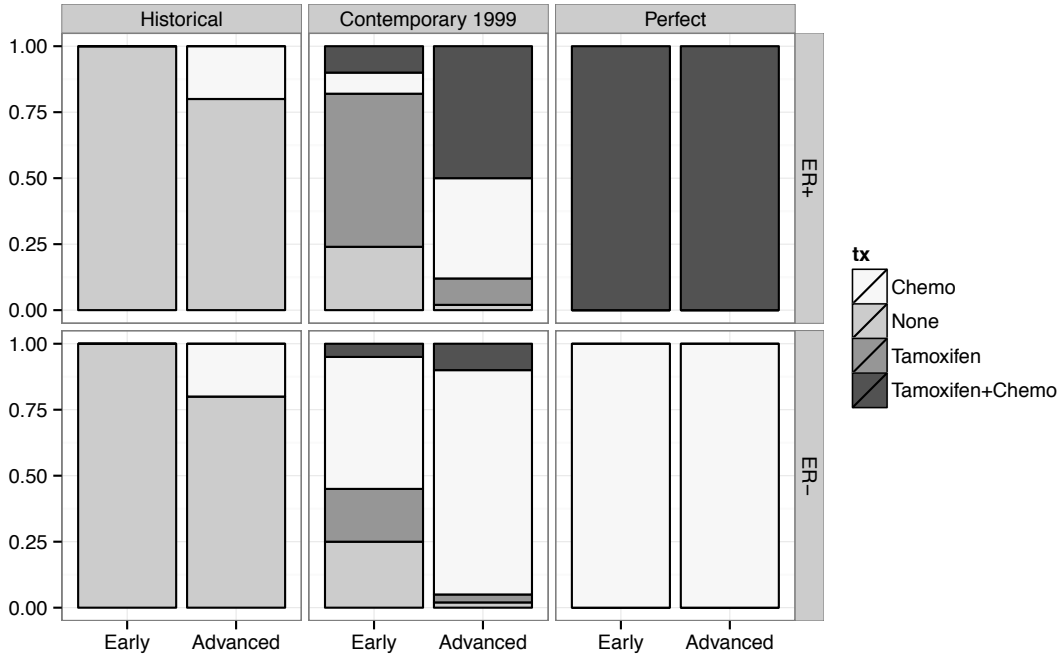


Figure 4.2

Cumulative incidence of breast cancer mortality in each arm of each trial, as a proportion of women alive at trial start. Mortality is reduced in the screening arm through a 15% decrease in advanced-stage disease. Conditional on stage and (contemporary trials only) receptor status at diagnosis, the same treatment is administered in both arms of the same trial, whereas the available treatments change across trials.

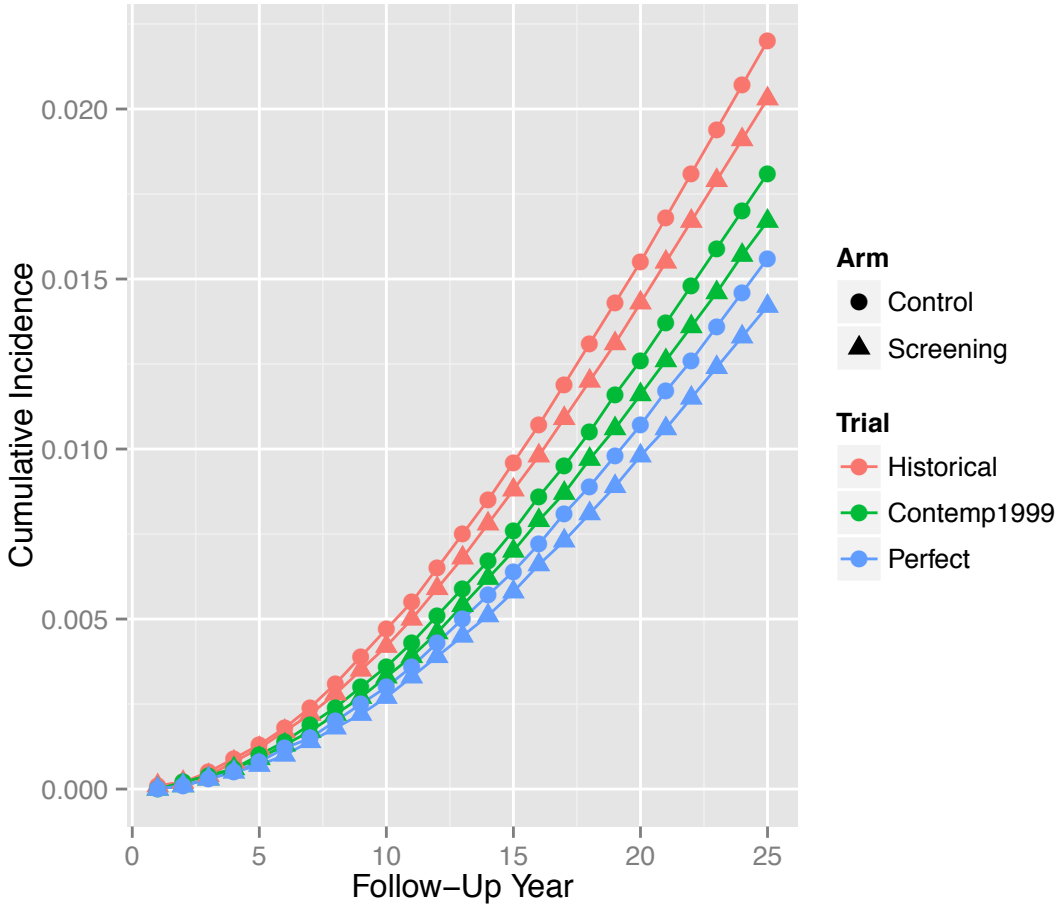
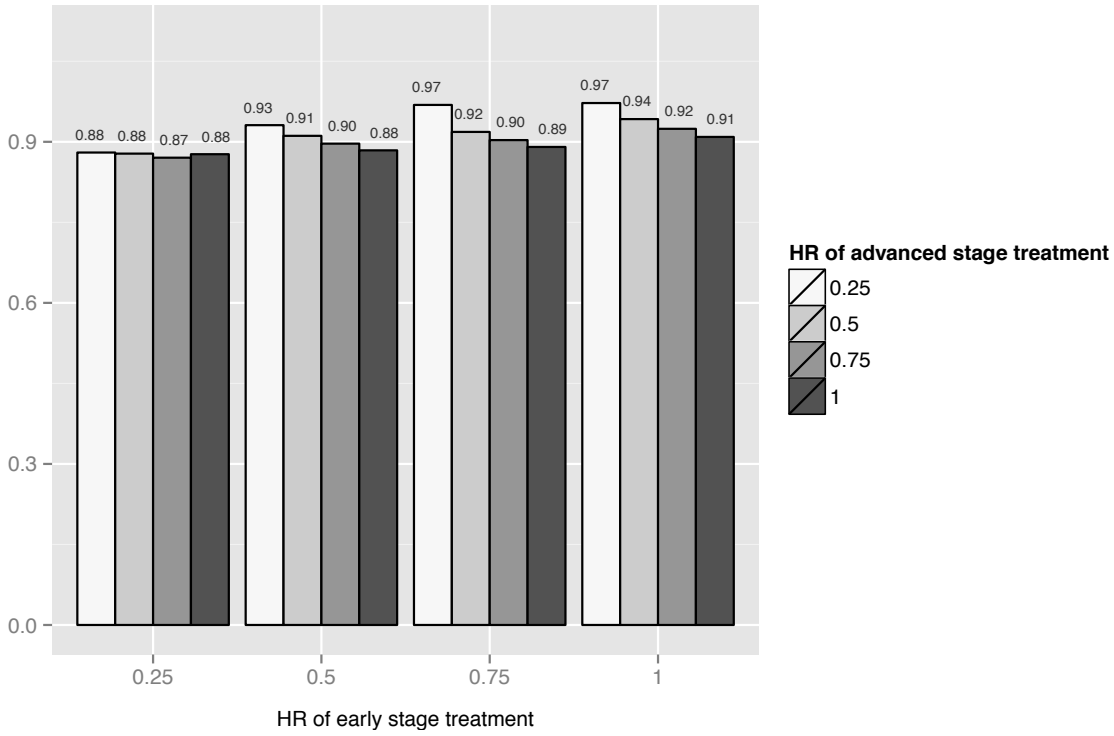


Figure 4.3

Mortality rate ratios for 16 hypothetical trials, each with a 15% stage shift due to screening and follow-up evaluated at 13 years. In each trial, there is a single treatment given to all early stage cases and another, single treatment given to all advanced cases. The trials differ in how effective the early and/or advanced stage treatments are in reducing breast cancer mortality. The x-axis indicates the hazard ratio or benefit of the early stage treatment for a given trial, while the bar colors indicate the hazard ratio for the advanced stage treatment. The right-most bar indicates that with no effective early or advanced stage treatments, the stage-shift model projects an MRR of 0.91 as the benefit of 15% screen-detection in an earlier stage.



Appendix

Table 4.A1

Data descriptions, sources and assumptions used to construct model parameters. Model parameter values are either specified in the table or a reference is given to the appropriate section of the manuscript.

Parameter	Data	Source	Assumptions	Values (Range)
Age at clinical diagnosis	Pre-screening era rates of clinical diagnosis	SEER, female diagnoses 1975-1979 by 5-year age groups (SEER 9, 2013 Nov submission)	Historical pre-screening era rates of clinical diagnosis approximate contemporary clinical incidence in the absence of screening. All major influences on diagnosis and survival since the 1970s are adequately captured by how we subsequently model screening and treatment	
Stage at clinical diagnosis	Pre-screening era stage distribution	SEER, female diagnoses 1975-1979 among ages 50-75 (SEER 18, 2013 Nov submission)	Historic stage "local" approximates early stage and "regional+distant" approximates advanced stage	**Local**: 49.6% **Reg/Dist**: 50.4% **Among Reg/Dist**: 82.9% Reg, 17.1% Dist
Receptor at clinical diagnosis	Frequency of ER/HER2 subtypes at diagnosis in 2010, conditional on stage	SEER, female diagnoses 2010 in ages 50-75 (SEER 18, 2013 Nov submission)	Stage-specific receptor frequencies in a mixed screen- and clinically-detected population apply to receptor frequencies in a clinically detected population	
Impact of screening on stage at diagnosis	Relative rate of advanced disease diagnosis across screening trials (Stage II+ or >20mm)	Autier 2009		Median 0.85 (0.69-0.97)

Parameter	Data	Source	Assumptions	Values (Range)
Baseline stage-specific mortality rate after clinical diagnosis	Pre-screening era cause-specific survival by local vs regional stage at diagnosis	SEER, female diagnoses 1975-1979 at ages 50-75 (SEER 18, 2013 submission)	Local approximates early stage Frequency-weighted regional/distant approximates advanced stage Exponential survival approximates true survival. Rate is an average of rates implied by observed survival at 5 and 10 years post-diagnosis Effect of age is negligible given effect of stage on survival No effect of receptor status on survival independent of stage and treatment	**Early**: 0.01992 **Advanced**: 0.10693
Impact of treatment on baseline mortality	EBCTCG meta-analyses of tamoxifen and polychemotherapy on disease-specific survival for operable breast cancer Cochrane review of effect of trastuzumab in early and metastatic breast cancer		Tamoxifen and chemo treatment effects are the same for early and advanced disease Treatment effect on overall survival approximates effect on disease-specific survival for trastuzumab	**Tamoxifen:** HR=0.70 for ER+, HR=1.0 for ER- **Chemo:** HR=0.775 for both **Tamoxifen+chemo:** HR=0.70*0.775 for ER+, HR=1.0*0.775 for ER- **Adding trastuzumab:** For HER2+ tumors, multiply above by 0.66 for early stage, and 0.82 for advanced stage

Table 4.A2

Impact of varying treatment on the cumulative incidence of breast cancer mortality observed in a screening trial.
Results reflect 13 years of follow-up in a female cohort of size 100,000 aged 50 in 2000, with 95% uncertainty intervals across 100 simulations. In all trials, screening reduces advanced disease by 25%.

	TREATMENT GIVEN IN THE SCREENING TRIAL		
	Historical Treatment	**1999 Treatment**	**Perfect Treatment**
Cumulative incidence, control group			
Control arm	0.0075 (0.0073,0.008)	0.0059 (0.0057,0.0064)	0.005 (0.0048,0.0054)
Screening arm	0.0064 (0.0062,0.0069)	0.005 (0.0049,0.0054)	0.0042 (0.004,0.0046)
Effect of screening, same treatment			
MRR	0.8533 (0.8479,0.8827)	0.8475 (0.8446,0.8865)	0.84 (0.8291,0.8724)
ARR	0.0011 (0.001,0.0013)	0.0009 (8e-04,0.001)	0.0008 (7e-04,0.001)
Effect of treatment, without screening*			
MRR		0.7867 (0.7797,0.8091)	0.6667 (0.6512,0.6962)
ARR		0.0016 (0.0015,0.0019)	0.0025 (0.0024,0.0028)
Effect of treatment, with screening^			
MRR		0.7812 (0.7785,0.8195)	0.6562 (0.635,0.6879)
ARR		0.0014 (0.0013,0.0016)	0.0022 (0.0021,0.0026)

* Compares control arm of the trial to control arm of the Historical Trial

^ Compares screening arm of the trial to screening arm of the Historical Trial

Table 4.A3

Impact of varying treatment on the cumulative incidence of breast cancer mortality observed in a screening trial, but with the early-stage mortality rate half halved from the main analysis.
Results reflect 13 years of follow-up in a female cohort of size 100,000 aged 50 in 2000, with 95% uncertainty intervals across 100 simulations. In all trials, screening reduces advanced disease by 15%, and the early-stage breast cancer mortality rate is reduced by half compared to the main analysis.

	TREATMENT GIVEN IN THE SCREENING TRIAL		
	Historical Treatment	**1999 Treatment**	**Perfect Treatment**
Cumulative incidence, control group			
Control arm	0.0067 (0.0066,0.0073)	0.0053 (0.0051,0.0058)	0.0045 (0.0044,0.005)
Screening arm	0.006 (0.0058,0.0065)	0.0047 (0.0046,0.0052)	0.004 (0.0039,0.0043)
Effect of screening, same treatment			
MRR	0.8955 (0.8793,0.9046)	0.8868 (0.8764,0.9083)	0.8889 (0.8693,0.9076)
ARR	0.0007 (7e-04,9e-04)	0.0006 (5e-04,7e-04)	0.0005 (5e-04,7e-04)
Effect of treatment, without screening*			
MRR		0.791 (0.7756,0.8205)	0.6716 (0.6557,0.7094)
ARR		0.0014 (0.0013,0.0017)	0.0022 (0.0021,0.0025)
Effect of treatment, with screening^			
MRR		0.7833 (0.7741,0.823)	0.6667 (0.6526,0.709)
ARR		0.0013 (0.0012,0.0015)	0.002 (0.0019,0.0023)

* Compares control arm of the trial to control arm of the Historical Trial

^ Compares screening arm of the trial to screening arm of the Historical Trial

Figure 4.A1

Mortality rate ratios for 16 hypothetical trials, each with a 25% stage shift due to screening and follow-up evaluated at 13 years. In each trial, there is a single treatment given to all early stage cases and another, single treatment given to all advanced cases. The trials differ in how effective the early and/or advanced stage treatments are in reducing breast cancer mortality. The x-axis indicates the hazard ratio or benefit of the early stage treatment for a given trial, while the bar colors indicate the hazard ratio for the advanced stage treatment. The right-most bar indicates that with no effective early or advanced stage treatments, the stage-shift model projects an MRR of 0.84 as the benefit of 25% screen-detection in an earlier stage.

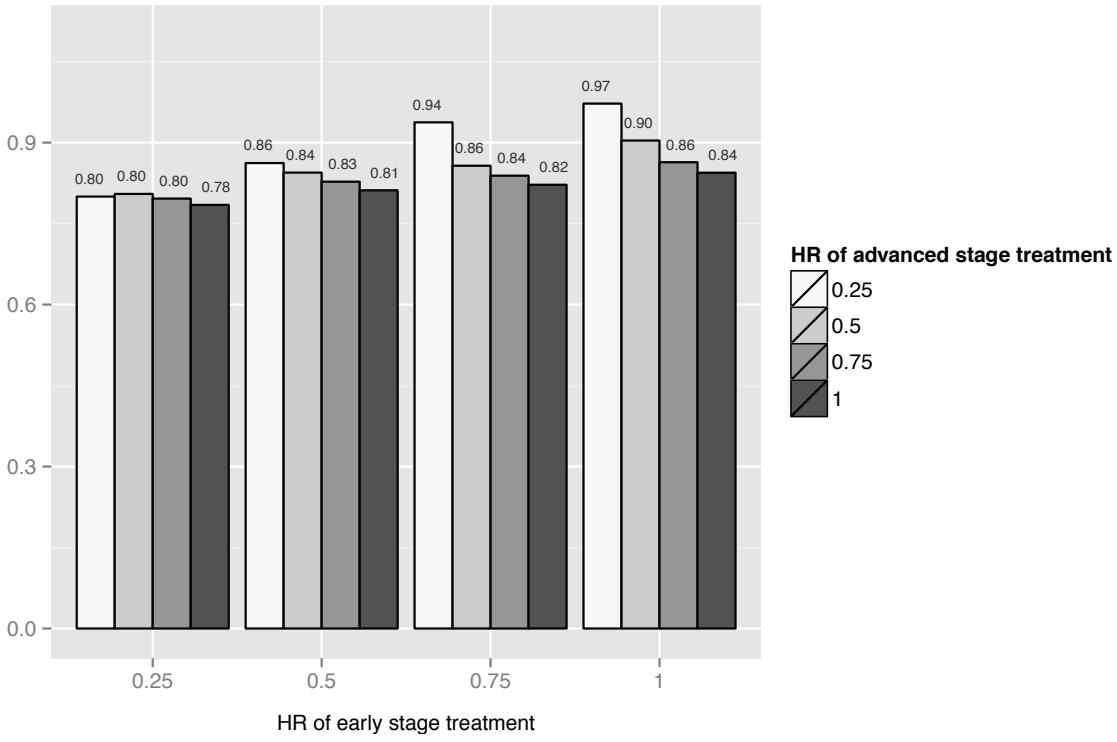
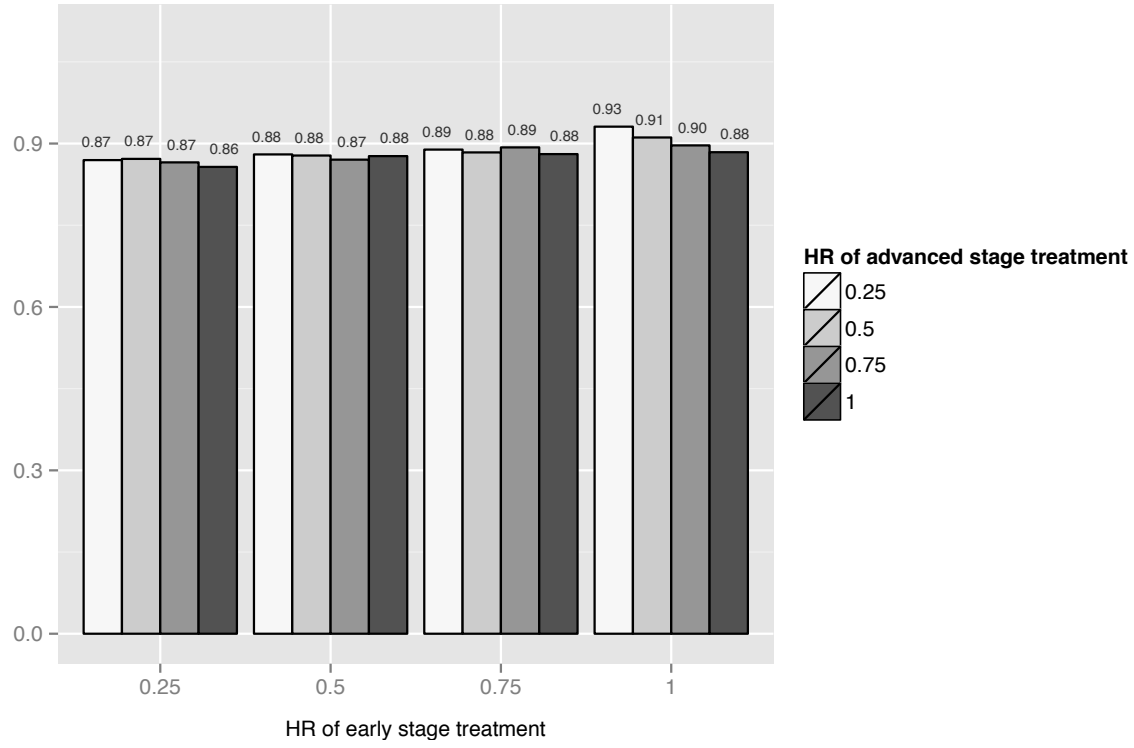


Figure 4.A2

Mortality rate ratios for 16 hypothetical trials, each with a 15% stage shift due to screening and follow-up evaluated at 13 years, but with the early-stage mortality rate half halved from the main analysis. In each trial, there is a single treatment given to all early stage cases and another, single treatment given to all advanced cases. The trials differ in how effective the early and/or advanced stage treatments are in reducing breast cancer mortality. The x-axis indicates the hazard ratio or benefit of the early stage treatment for a given trial, while the bar colors indicate the hazard ratio for the advanced stage treatment. The right-most bar indicates that with no effective early or advanced stage treatments, the stage-shift model projects an MRR of 0.88 as the benefit of 15% screen-detection in an earlier stage.



CHAPTER 5: CONCLUSION

To keep pace with the demand for research evidence to support health policy, comparative effectiveness research (CER) must embrace not only new forms of empirical research designs but also alternative methods for evidence generation, including modeling. While modeling is often dismissed as providing weak guidance at best,[1,2] there are policy questions in cancer screening and diagnosis that simply cannot be fully supported by the results of empirical studies. The mortality endpoint is too important to evaluating cancer interventions to circumvent using more easily acquired endpoints or qualitative speculation. This dissertation thus centers on the premise that modeling can enhance CER in cancer by providing quantitative mortality projections for questions that are only incompletely answered by empirical comparative effectiveness data.

The first dissertation model advances cancer CER by providing an easily re-usable model for evaluating new diagnostic tests in cancer, along with novel projections of how clinical use of the 21-gene recurrence score impacts breast cancer mortality, quality of life and cost. Chapter 2 outlines the model structure and its various options that accommodate a range of applications, and additionally provides two "case studies" of the 21-gene recurrence score. Both case studies demonstrate the flexibility of the model to different forms of input CE data, and the first case study also provides a cross-validation of the model. Given the lack of empirical data for validation, our close replication of the results of an existing model provides some evidence that our model produces reasonable projections. Our second case study presents the first translation of the real-world clinical impact of diagnostic testing on treatment patterns to mortality, quality of life and costs. The results suggest that while the mortality impact of using the 21-gene recurrence score may be minimal in clinical practice, the test's effect of decreasing chemotherapy use improves quality of life and offsets the cost of the assay. Any interested user

can download our Windows application and follow the video and written tutorials to replicate both case studies, test them using different input parameters, or develop their own model.

The second dissertation model contributes to cancer CER both a template for much-needed early evaluations of new cancer screening biomarkers as well as specific projections for the long-term benefits and harms of incorporating the new biomarker PCA3 into prostate cancer screening. Chapter 3 steps readers through the three major modeling components needed to evaluate a screening biomarker: a natural history model of cancer progression, early-stage data on the biomarker, and a sub-model of the survival benefit of early detection. Each component is accompanied by the corresponding components of our model of PCA3. In that application, we introduce a novel sub-model of the survival benefit of early detection that reflects the intuition that earliness of detection is related to the likelihood of cure. Our results highlight that the screening strategies that most effectively use PCA3 to decrease harms will also decrease benefits, but we do identify a strategy that would substantially lower false positives and overdiagnosis while preserving 85% of lives saved by PSA-only screening.

The third model offers cancer CER a general framework for evaluating how advances in cancer treatments may alter the mortality benefit of screening and updates the mortality impact of mammography screening to account for contemporary breast cancer treatments. Chapter 4 introduces the stage-shift model of screening and shows how the mortality impact of screening can be decomposed into a stage shift and stage-specific survival after clinical diagnosis. Survival depends on the treatment received. This simple model structure is used to conduct virtual screening trials with identical stage shifts but different treatment distributions. We find that in the case of breast cancer, contemporary treatments do not substantively alter the mortality impact of mammography screening. While this result runs counter to prevailing hypotheses, it is consistent with the similar effectiveness across stages of contemporary treatments. Our sensitivity analyses show that trial results would change if treatment efficacies

by stage were different, but the underlying mortality rates set boundaries on how much treatments can impact screening for a given stage shift.

A unique strength of all three studies is that the models are naturally primed for future research. The first model can evaluate any new diagnostic test in cancer given data on how testing influences treatment distributions and mortality after treatment. New applications may highlight useful extensions to the existing model, but the primary structure should be highly transferrable. In particular, head-to-head comparisons of diagnostic tests may be valuable as competing assays continue emerge. Personalized medicine is a hot topic and commercial interest in genomic test development is strong.[3–6] There are several alternatives to the 21-gene recurrence score for assessing breast cancer prognosis and the need for chemotherapy, for example.[7] The first model is well-suited to respond to CER needs as the number of diagnostic test candidates expands. In addition, the recent advent of the R-to-web extension "RShiny" (www.rshiny.com) makes it feasible to build a web-based, platform-independent version of the user interface that would further increase access to the model.

The procedure for modeling the impact of new screening biomarkers used in the second model can similarly apply to other screening biomarkers. The PCA3 model itself could also be re-applied to evaluate a different prostate cancer biomarker, if data were available on biomarker levels in cancer cases and non-cases and the biomarker's correlation with PSA. One possibility biomarker is TMPRSS2:erg, another urinary assay that, like PCA3, has shown promise in early-stage studies.[8,9] More immediately, longitudinal PCA3 data currently being collected by the EDRN could be used to update the assumptions made regarding PCA3 trajectories over time.

The third model is ripe for investigating additional situations in which differential treatment has obscured the impact of screening. We used the model to evaluate treatment advances, but it could also be used to investigate treatment disparities. In prostate cancer, disease incidence and mortality has historically been worse in blacks than whites, and

treatments have disseminated into the population differentially by race.[10,11] As PSA screening increased in the 1990s, the use of curative treatments increased among both races but remained higher among whites.[12,13] Our model could compare a virtual screening trial of whites to a trial of blacks. Both trials would have the same screening stage shift, but with greater incidence and less curative treatment lower in the black trial. If the results showed greater benefit of screening in blacks than whites, then the recent controversial USPSTF recommendation against PSA screening would actually increase disparities in prostate cancer mortality. This work could highlight a need for race-specific screening guidelines.

As discussed in depth in Chapters 2-4, all analyses using these models will suffer from the limitations inherent to modeling. Models simplify reality. All three studies involve numerous assumptions that make the models feasible to construct in a fairly transparent manner. Sound scientific logic, calibration and cross-validation were used to strengthen the models, but there is no certainty that the structures used to represent latent, largely unknown disease processes are correct. However, the goal in modeling is to be correct *enough*, given existing scientific knowledge, for the resulting evidence to be useful. All three models were designed with this aim in mind.

This dissertation provides modeling tools and evidence for policy settings in which empirical data fall short. Cancer technology development will continue to outpace the observation of cancer mortality. Modeling can address this limitation of timing inherent in studies by synthesizing available evidence into a more complete picture of intervention effectiveness. This dissertation attempts to provide models that enhance the capability of CER to support high-quality policy decisions in cancer.

References

1. Bach, P. B. Raising the Bar for the U.S. Preventive Services Task Force. *Ann. Intern. Med.* **160,** 365–366 (2014).
2. Melnikow, J., LeFevre, M., Wilt, T. J. & Moyer, V. A. Counterpoint: Randomized trials provide the strongest evidence for clinical guidelines: The US Preventive Services Task Force and Prostate Cancer Screening. *Med. Care* **51,** 301–303 (2013).
3. Arteaga, C. L. & Baselga, J. Impact of Genomics on Personalized Cancer Medicine. *Clin. Cancer Res.* **18,** 612–618 (2012).
4. Bates, S. Progress towards personalized medicine. *Drug Discov. Today* **15,** 115–120 (2010).
5. Kalia, M. Personalized oncology: Recent advances and future challenges. *Metabolism* **62, Supplement 1,** S11–S14 (2013).
6. Roukos, D. H., Murray, S. & Briasoulis, E. Molecular genetic tools shape a roadmap towards a more accurate prognostic prediction and personalized management of cancer. *Cancer Biol. Ther.* **6,** 308–312 (2007).
7. Harbeck, N., Sotlar, K., Wuerstlein, R. & Doisneau-Sixou, S. Molecular and protein markers for clinical decision making in breast cancer: today and tomorrow. *Cancer Treat. Rev.* **40,** 434–444 (2014).
8. Leyten, G. H. J. M. *et al.* Prospective multicentre evaluation of PCA3 and TMPRSS2-ERG gene fusions as diagnostic and prognostic urinary biomarkers for prostate cancer. *Eur. Urol.* **65,** 534–542 (2014).
9. Wei, J. T. Urinary biomarkers for prostate cancer. *Curr. Opin. Urol.* **25,** 77–82 (2015).
10. Chu, K. C., Tarone, R. E. & Freeman, H. P. Trends in prostate cancer mortality among black men and white men in the United States. *Cancer* **97,** 1507–1516 (2003).
11. Hankey, B. F. *et al.* Cancer surveillance series: interpreting trends in prostate cancer--part I: Evidence of the effects of screening in recent prostate cancer incidence, mortality, and survival rates. *J. Natl. Cancer Inst.* **91,** 1017–1024 (1999).
12. Bolla, M. *et al.* Improved survival in patients with locally advanced prostate cancer treated with radiotherapy and goserelin. *N. Engl. J. Med.* **337,** 295–300 (1997).
13. Bill-Axelson, A. *et al.* Radical prostatectomy versus watchful waiting in early prostate cancer. *N. Engl. J. Med.* **364,** 1708–1717 (2011).