



FLEXKB DOCUMENT
Version: HI.001.10282008.65053
Document generated: 10/28/2008

UNIVERSITY OF MICHIGAN

Important note: This document will be updated periodically. The most current version is available at <http://cisnet.cancer.gov/profiles>. Note that unlike most PDF documents, the CISNET model profiles are not suitable for printing as they are not typically written or read in sequential fashion.

We recommend you let your interests guide you through this document, using the navigation tree as a general guide to the content available.

The intent of this document is to provide the interested reader with insight into ongoing research. Model parameters, structure, and results contained herein should be considered representative but preliminary in nature.

We encourage interested readers to contact the contributors for further information.

Go directly to the: [Reader's Guide](#).



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)



READERS GUIDE

Core Profile Documentation

These topics will provide an overview of the model without the burden of detail. Each can be read in about 5-10 minutes. Each contains links to more detailed information if required.

Model Purpose

This document describes the primary purpose of the model.

Model Overview

This document describes the primary aims and general purposes of this modeling effort.

Assumption Overview

An overview of the basic assumptions inherent in this model.

Parameter Overview

Describes the basic parameter set used to inform the model, more detailed information is available for each specific parameter.

Component Overview

A description of the basic computational building blocks (components) of the model.

Output Overview

Definitions and methodologies for the basic model outputs.

Results Overview

A guide to the results obtained from the model.

Key References

A list of references used in the development of the model.

Further Reading

These topics will provide a intermediate level view of the model. Consider these documents if you are interested gaining a working knowledge of the model, its inputs and outputs.

Advanced Reading

These topics denote more detailed documentation about specific and important aspects of the model structure



MODEL PURPOSE

SUMMARY

This page summarizes the purposes for which this model was developed.

PURPOSE

The model provides a quantitative link between dissemination of cancer control processes and their impact on population and public health measures of cancer incidence, survival and mortality. Its purpose is to unravel the myriad causes and relationships that underlie recent trends in prostate cancer incidence and mortality, to quantify the relationships in terms of model parameters, and to enable researchers to perform inference on these parameters by means of confidence intervals and hypothesis tests.

The model provides tools by which national population and cancer registry data may be analyzed, so that the population impact of cancer control processes may be understood and predicted. It exists that researchers might predict short- and long-term trends in national incidence and mortality under various scenarios; might analyze racial disparities as they pertain to factors associated with trends in treatment, survival, incidence and mortality; and might determine and evaluate optimal screening strategies.

A particular goal of this model is to enable researchers to determine in what way, if at all, PSA screening of asymptomatic men is linked to the recent decline in prostate cancer mortality. In fact, the model has already generated predictions for prostate cancer incidence and mortality under current PSA utilization patterns, and for the baseline case of no PSA screening. The latter prediction is counterfactual, in that it expresses what incidence and mortality would have been during the years 1970-2000 if there had been no PSA screening, other things being equal. Thus it yields an estimate of the differences in incidence and mortality that are purely associated with PSA utilization.



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)



MODEL OVERVIEW

SUMMARY

This document provides an overview of the modeling effort, and describes the model itself in general terms.

PURPOSE

This is a model of prostate cancer incidence and mortality. It was developed to analyze national population and cancer registry data. It is used to understand, predict and optimize the population impact of cancer control processes in prostate cancer. See [Model Purpose](#) for more details.

BACKGROUND

Excluding skin cancers, prostate cancer is the most common cancer in American men. It claims over 40,000 lives annually, ten percent of cancer deaths among men, and is second only to lung cancer as a cause of cancer deaths¹. Progressive prostate cancer is a serious disease. Thousands of men suffer pain and complications and die prematurely from progressing tumors.

Management and control of prostate cancer is a significant public health problem. For more than a decade since the introduction of PSA testing in the late 80s, the incidence rates of newly diagnosed prostate cancers have seen a dramatic increase to over 190,000 cases in early 90s, followed by an equally dramatic decline² ([Figure_1](#)). At the same time, mortality slowly increased from the 70s to the early 90s, and has been declining since then³ PSA screening has spread through the population because of the hope that it ultimately may reduce mortality. But the mere fact that screens can detect organ-confined prostate cancer does not in itself constitute a sufficient ground for their implementation. Screening cannot be justified unless patients who are screened actually have improved outcomes, and this has not yet been shown.

To make appropriate decisions regarding treatment and public health policy, we must understand the causes of these trends. To do so, models are needed that can unravel and disentangle all the factors behind the observed population trends, including length bias, overdiagnosis, early detection of cases that would become clinical, shifts in stage and grade of cancer associated with early detection, and other possible factors, such as a change in the survival curve within stage and grade following the advent of PSA screening.

The model reported here has several distinctive features that meet this challenge. These include:

- Estimation from population data instead of from screening trial data.
- A flexible regression framework, accomodating explicit adjustments for differences in screening and treatment utilization patterns.
- Analytic, rather than simulation-based, procedures for estimation and prediction.
- The identification of within-stage shift as a factor affecting survival and mortality.



- Confidence intervals and tests of statistical hypotheses for all model [parameters](#).
- A general structure, making the model applicable to cancer sites other than the prostate.

Before describing the "guts" of the model, we briefly discuss these features.

Estimation from population data.

The [parameters](#) in our model are estimated from population databases such as SEER, not from screening trial data. This mode of estimation is possible because we incorporate random PSA schedules into the estimation procedure. This approach is appropriate, since the focus of CISNET is on population trends in incidence and mortality. To measure such trends, a model must unravel and disentangle a set of competing risks and confounded effects, including length bias, overdiagnosis, advancement of diagnosis due to screening, stage shift due to early detection, and other possible effects, such as within-stage shift. No randomized trial for which data are currently available has been designed to measure all these effects. To model such effects, it is critical that methods be developed that can exploit population data for estimation. Our model does this.

Since our estimation procedure is based on population databases, we are able to exploit the wealth of information available from these databases. The consequences for precision and power are substantial, since population databases are typically much larger than the number of individuals participating in a screening trial. Increased power enables us to obtain reliable estimates on a potentially larger set of model [parameters](#) than would be possible were the estimation based on screening trials.

A flexible regression framework.

The model is constructed within a flexible framework that uses concepts of survival analysis, yet is not limited to the standard Cox proportional hazard model. This framework accommodates adjustments for variable screening and treatment utilization patterns through built-in lead-time, length-bias and stage- and within-stage shift, all of which may affect survival. This framework allows the researcher to derive realistic estimates of mortality through joint modeling of incidence and survival in a dynamic population environment. For example, clinical covariates being equal, the model generally would provide different survival estimates for subjects from low and high PSA utilization areas.

Since this regression framework is applied to population data, there is no need to perform additional calibration of model predictions after [parameter](#) estimation. All adjustments of the model are explicitly built into the fitting procedure. In case the fit is unsatisfactory, this framework naturally leads the researcher or policymaker to directly examine model assumptions and [parameter](#) values. Thus the framework is flexible, allowing reviewers, researchers outside the development team, and policymakers to analyze and evaluate the model, creating an environment conducive to further model improvement.

Analytic procedures for estimation and prediction.

A second distinctive feature of our model is that it is analytic in all its components. We do not use simulation or stochastic approximation either to fit the model or to predict



from it. This fact has two consequences, one pertaining to ease of model fitting and prediction, one pertaining to model interpretation.

A fit or prediction run for our model takes a few minutes on a standard PC, much less time than if we relied on simulation. This results in a quick feedback and consequently allows interactive dialog between the model and the user. This feature is particularly important for providing a policy maker with a tool to quickly evaluate a number of cancer control hypotheses in an interactive environment. We have already developed a working prototype software package (SCANS, Self-Consistency Analysis of Surveillance) for Windows.

In addition to the practical advantage of speed, the analytic nature of our model means that it is transparent. The [parameters](#) can be directly interpreted in terms of processes of interest, so that the model itself is in no way a "black box."

Within-stage shift.

It is customary to explain survival and mortality differences associated with PSA screening by shifts in stage and grade of the cancer associated with the lead time between screen diagnosis and the time at which an individual would have been diagnosed clinically. Under the current model, however, we find evidence of a survival shift within stage, associated with PSA dissemination. For more information, see [Within Stage Shift](#).

Additional distinctive features of this model.

We provide confidence intervals and tests of statistical hypotheses for all estimated model [parameters](#). Thus, the user has an idea as to the significance of model findings.

While the model is applied to prostate cancer, its structure is general and open to immediate application to other diseases.

MODEL DESCRIPTION

The model provides a means by which [parameters](#) may be estimated that enable us to explain and predict trends in incidence, survival and mortality. As suggested in the Background section above, the model is probabilistic -- yielding p-values and confidence intervals -- and is accompanied by procedures that permit estimation from the same kind of data that we are seeking to predict and explain, namely, large population databases.

Model Assumptions

Please see [Assumption Overview](#) for the assumptions on which the model is based.

Model Inputs

Please see [Parameter Overview](#) for a list of model inputs.

Model Outputs

The model yields estimates of a set of [parameters](#) that together constitute a comprehensive model for prostate cancer incidence, survival, and mortality.



These [parameters](#) govern such basic characteristics of the model as age at tumor onset, sojourn time, lead time, overdiagnosis, delay time, sensitivity of the PSA test, and the correlation between the age at tumor onset and the sojourn time. From these basic characteristics follow estimates of incidence a function of calendar year, age, stage, and grade; survival as a function of stage, grade, and delay time; and mortality by calendar year and age. From these [parameters](#) we finally derive estimates of the effect of PSA screening on prostate cancer incidence and mortality.

For more information, see [Output Overview](#).

Model Limitations

Since the data are observational, we do not have the benefit of complete elimination of confounders, as is possible in a well-conducted clinical trial. The urgent need to gain understanding of the processes, however, and the long follow-up time required by a screening trial, do not afford us the luxury of waiting for experimental results.

Because the model is based on past data, its generalizability to the future may be limited.

The current version of the model does not explicitly describe PSA growth.

CONTRIBUTORS

We gratefully acknowledge the collaboration of the following individuals.

Dr. Ray Merrill of the Department of Health Sciences, Brigham Young University, helped in model building, interpretation and prediction of national prostate cancer trends.

Dr. Marco Zaider, Head of Brachytherapy at the Memorial Sloan-Kettering Cancer Center, brought his expertise in prostate cancer treatment to the project. He assisted us in analyzing and interpreting clinical data, providing a link between screening strategies and prostate cancer post-treatment survival.

Dr. Gilda Garibotti worked on computer implementation of the profile information matrix methodology in the survival analysis module. She provided advice on software development and implementation of survival analysis machinery in the population model.

Dr. Aniko Szabo provided help on software development and implementation of methods and population models. She especially provided advice on the integration of the model software into the population software shell, and helped in testing computer code that implements extended population models.

REFERENCES:

- ¹ Stewart, S.L., King, J.B., Thompson, T.D., Friedman, C., Wingo, P.A. "Cancer mortality surveillance--United States, 1990-2000" in MMWR Surveillance Summary 2004; 53: 3: 1-108
- ² Potosky, A.L., Feuer, E.J., Levin, D.L. "Impact of screening on incidence and mortality of prostate cancer in the United States" in Epidemiological Review 2001; 23: 1: 181-186



University of Michigan
Model Overview
References:

³ Hsing, A.W., Devesa, S.S. "Trends and patterns of prostate cancer: What do they suggest?" in Epidemiologic Reviews 2001; 23: : 3-13



ASSUMPTION OVERVIEW

SUMMARY

In this section we summarize the main assumptions on which the model is based.

BACKGROUND

Researchers generally agree that prostate cancer is the result of an irreversible transition of the disease through three consecutive stages: the disease free stage, the pre-clinical stage and the clinical stage. This three-stage model entails the following potential time points in an individual's life: birth, [onset](#) of prostate cancer, time of clinical diagnosis, time of death due to prostate cancer, and alternatively the time of death due to a competing risk.

Although this model may be accurate as far as it goes, it does not capture the processes currently affecting incidence, survival and mortality. The dissemination of [TURP](#) and, more drastically, the dissemination of PSA testing have made the picture more complicated, because both [TURP](#) and the PSA test can advance the diagnosis of prostate cancer.

In addition, it has customarily been believed that stage shift is the only reasonable explanation for any benefit derived from early detection. We do not make this assumption, and in fact have found evidence to the contrary (see [Introduce Within Stage Shift](#)).

For both these reasons, a more complex set of assumptions must be spelled out.

ASSUMPTION LISTING

Tumor [onset](#) (See [Age At Tumor Onset](#) for details.)

- The baseline hazard of tumor onset may depend on age.
- The hazard of tumor onset may depend on calendar year.
- The effect of calendar year on the hazard of tumor onset is multiplicative.

[Sojourn Time](#) (See [Sojourn Time Distribution](#) for details.)

- Baseline [Sojourn Time](#) may depend on age.
- The hazard function associated with baseline [Sojourn Time](#) may include a multiplicative trend in calendar time.
- [Sojourn Time](#) may depend on age at tumor onset.
- Given the time of tumor onset, [Sojourn Time](#) does not depend on the cancer screening process.

[Delay Time](#). The distribution of the duration of the [latent_disease_stage](#) is an average over random patterns of all possible modes by which the disease can be detected. (See [Incidence Model](#) for details.) These include

- Clinical diagnosis through symptoms



- PSA screening
- Transurethral resection of the prostate ([TURP](#)).

Screening tests occur randomly in time, subject to the following assumptions (see [PSAScreening Model](#) for details):

- Age at first PSA test has a distribution (alternatively, hazard function) that depends both on age and calendar time.
- The times between consecutive PSA tests occur as a non-homogeneous Poisson process, with an intensity that depends on age and calendar year.
- The sensitivity of the screening test is an increasing function of time since tumor onset.

Survival time after diagnosis follows a semiparametric regression model. (See [Survival Component](#) for details.)

Mortality in the population can be adequately modeled by combining information from the incidence and survival models. (See [Mortality Model](#) for details.)

Within-Stage Differences in Prognosis. We allow the possibility that stage and grade at diagnosis are not the only variables associated with a patient's prognosis for survival. (See [Introduce Within Stage Shift](#).)



PARAMETER OVERVIEW

SUMMARY

This document lists and defines the inputs to the modeling algorithm.

BACKGROUND

The parameters discussed on this page are parameters of a **model**, not of a population. Thus, in this context, the term "parameter" has an entirely different meaning from the classical statistical use of the term.

In classical statistics, a parameter is a number or set of numbers that characterize a population. Typically, we estimate parameters by drawing a sample from the population, measuring a variable or set of variables on each individual in that sample, and using these measurements (i.e., using a set of data) as input to an estimation algorithm. The estimation algorithm produces parameter estimates as output. These may be point estimates, interval estimates, p-values, or higher-dimensional objects such as densities or cumulative distribution functions.

In the current context, however, "parameter" does not refer to a characteristic of the population but rather to any input to the modeling algorithm. Thus, even a set of measurements made on each individual in a sample drawn from the population are considered "parameters" if they are used as input to the modeling algorithm.

PARAMETER LISTING OVERVIEW

The inputs to the model (also called the "model parameters," as explained in the Background section of this page) consist of population data as well as "given" parameters and distributions provided by the National Cancer Institute (NCI). These include:

- The distribution of PSA utilization in the population. This distribution is based on an algorithm that can be used to simulate life histories of the times that individual men undergo PSA tests. For more information see [base Case PSA](#).
- Surveillance, Epidemiology and End Results (SEER) data on every individual diagnosed with prostate cancer in nine areas of the United States (San Francisco-Oakland, Connecticut, Detroit, Hawaii, Iowa, New Mexico, Seattle, Utah, and Atlanta), more than 350,000 cases ([SEER_Medicare](#)). The data include tumor characteristics as well as standard follow-up and outcome variables. In particular, for each age a (over 50) and each year t the number of new prostate cancer cases $C(a, t)$ is derived and "fed" to the model. For details see [Likelihood In The Incidence Model](#).
- Population count files belonging to the same areas from which the prostate cancer case data were obtained. From this source the number of people at risk for prostate cancer for each age a and each year t is derived, $P(a, t)$, and "fed" into the model. For details see [Likelihood In The Incidence Model](#).
- Age distribution in the U.S. population in the year 2000 for men over 50.
- Risk of death from other causes, derived from the [Human Mortality Database](#)¹.





University of Michigan
Parameter Overview
References:

REFERENCES:

- ¹ Wilmoth, John R. (Director), Shkolnikov, Vladimir Shkolnikov (Co-Director),
"Human Mortality Database (HMD)." 2003;



COMPONENT OVERVIEW

SUMMARY

This document outlines the analytic components of which the model is constructed.

OVERVIEW

The model is composed of three distinct components by which predictions and estimates are made based on population data.

The [Incidence Component](#) takes population data as input and yields estimates and predictions of prostate cancer incidence by calendar year and age. In addition it yields predictions both in the presence and in the absence of PSA testing, thereby yielding an estimate of the difference in prostate cancer incidence that is due to the presence of PSA testing.

The [Survival Component](#) also takes population data as input. It yields a model for the relationship between a set of covariates (including age, year of diagnosis, cancer stage and tumor grade) and a man's survival prognosis.

The [Mortality Component](#) combines the [Incidence Component](#) and the [Survival Component](#). It yields estimates and predictions of prostate cancer mortality by calendar year, age, and presence or absence of PSA testing. Thus, similar to the [Incidence Component](#), the [Mortality Component](#) yields an estimate of the difference in prostate cancer mortality that is due to the presence of PSA testing.

COMPONENT LISTING

[Incidence Component](#)

[Survival Component](#)

[Mortality Component](#)



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)



OUTPUT OVERVIEW

SUMMARY

This page lists and describes the statistics computed by the model, the [parameters](#) for which these statistics serve as estimates, and the predictions available from the model.

OVERVIEW

The model yields estimates of a set of [parameters](#) (some one-dimensional, some multi- or high-dimensional) that together constitute a comprehensive model for prostate cancer incidence, survival, and mortality.

OUTPUT LISTING

- Prostate cancer incidence as a function of calendar year, age, and presence or absence of PSA testing. (See [Incidence Figure](#).)
- Survival as a function of calendar year, age, stage, grade, and screening schedule. (See [Survival Component](#).)
- Mortality by calendar year, age, and presence or absence of PSA testing. (See [Mortality Component](#).)
- Mean lead time as a function of birth cohort. (See [define Lead Time](#) and [Results Overdiagnosis Lead Time](#).)
- Overdiagnosis as a function of birth cohort. (See [define Overdiagnosis](#) and [Results Overdiagnosis Lead Time](#).)
- Delay time as a function of calendar year and age, with and without PSA testing. (The latter is a counterfactual scenario; see [define Delay Time](#) and [Mean Posterior Delay Time Marginal Incidence Model](#).)
- Relationship between delay time and survival. (See [Within Stage Shift By Delay Time](#).) Note that this estimate entails differences in survival (or prognosis) associated with differences in screening schedule, even after adjustment has been made for stage and grade of cancer. We call this phenomenon within-stage shift. To our knowledge this phenomenon has not been discovered or quantified by any other research group.
- Estimates of the differences in incidence and mortality that are due to PSA screening. These estimates are based on a scenario in which incidence and mortality are estimated in the counterfactual case of no PSA screening. (See [results Age Adjusted Incidence Mortality](#).)
- Parameters governing the distribution of baseline sojourn time. (See [Incidence Model](#), [Analysis Of Population Data](#) and [Table_1](#).)
- Parameters governing age at tumor onset. (See [Incidence Model](#), [Analysis Of Population Data](#), and [Table_1](#).)
- Sensitivity of the PSA test. This was estimated at 100% by the model, but can be set to zero to predict what incidence would be in the absence of PSA. (See [Modeling Cancer Detection Through Screening](#) and [Analysis Of Population Data](#).)
- Correlation between age at onset and sojourn time. (This was found to be negligible, and removed from the model; see [Analysis Of Population Data](#).)



RESULTS OVERVIEW

SUMMARY

This page lists discoveries that have been made through the current modeling effort.

OVERVIEW

The model constitutes a framework for analysis of population databases. Within this framework, and by means of its estimation procedures, researchers obtain estimates of parameters that matter in the quest to understand the causes and processes of change in prostate cancer incidence, survival, and mortality. These parameters represent factors that either cause, or are associated with, differences in the outcomes that researchers and the public ultimately care about: survival prognosis for those diagnosed with prostate cancer, and mortality in the population due to prostate cancer.

RESULTS LIST

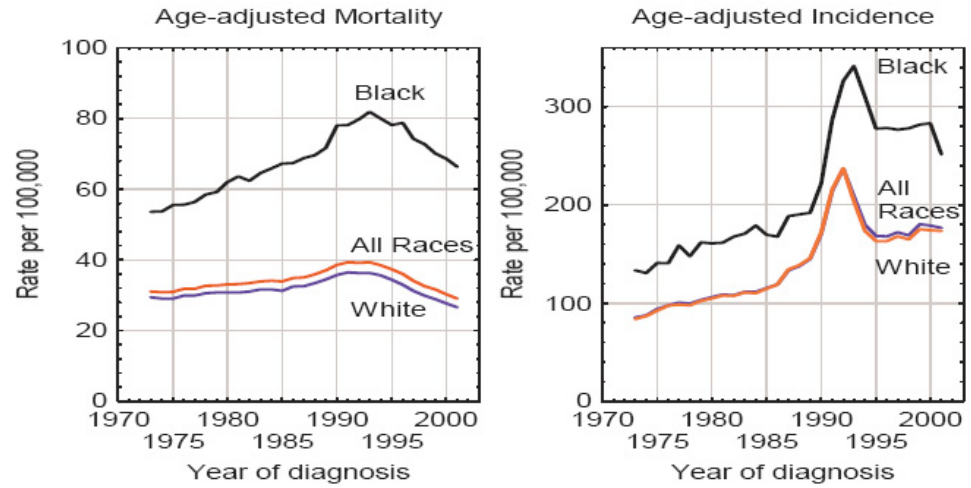
- [resultsAgeAdjustedIncidenceMortality](#), and consequently
- Estimation of the difference in incidence that is due to PSA testing.
- [resultsAgeAdjustedIncidenceMortality](#), and consequently
- Estimation of the difference in mortality that is due to PSA testing.
For more information on results regarding incidence and mortality, see [results Age Adjusted Incidence Mortality](#).
- Estimation of [Delay Time](#) without using survival data. This represents an independent significant prognostic factor for post-treatment survival, particularly for cancers in the localized-regional stage. See [Mean Posterior Delay Time Marginal Incidence Model](#).
- Identification of differences in survival associated with differences in [Delay Time](#). The changes in survival associated with early detection have been customarily modeled by stage shift. We have found, however, that, even after stage has been accounted for, differences in delay time (the time from onset of cancer until detection) are associated with differences in survival. See [Within Stage Shift By Delay Time](#).
- [Publications](#) accepted or submitted (click on the link for a list).



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)



FIGURE_1



Figure_1. Prostate cancer incidence and mortality rates by year of diagnosis age-adjusted to US population in year 2000. Data from the Surveillance, Epidemiology and End Results (SEER) database, National Cancer Institute.



University of Michigan
parameter



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)

PARAMETER

In this model profile, the term "parameter" is used in the classical statistical sense. In this sense a parameter is a number or set of numbers that characterize a population. A primary object of classical statistics is to use data, also referred to as a sample, to estimate the parameters of a population. This sense of the term is entirely different from the sense referred to by the phrase [Parameter Overview](#). Please see the Background section of that document for an explanation of the difference between the two senses of the word.

For information on the particular parameters estimated by the model, see [Output Overview](#).



University of Michigan
Within Stage Shift



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)

WITHIN STAGE SHIFT

The effects of [over diagnosis](#), [Length Bias](#) and [Lead Time](#) result in remarkable changes in the meaning of clinical covariates at diagnosis. With the introduction of screening, however, the prognostic value of such covariates is modified. The prognosis for cases diagnosed in the screening era is markedly different as compared to cancer cases from unscreened populations. This effect remains unexplained even when survival is adjusted for stage and grade ([Figure_2](#)). As PSA screening is intensified with the dissemination of the test in the U.S. population, survival in localized stage is improving while survival in distant stage is worsening as discussed above. A similar effect in the localized stage might be found before PSA was introduced, in association with early detection through TURPs. In contrast to the situation in a clinical trial, straightforward conditioning on clinical covariates in the analysis of population data may be misleading, and special care is needed to adjust for screening patterns in the population. For the same reason, the results of clinical trials are not immediately generalizable to the population setting.



University of Michigan
define Onset

DEFINE ONSET

Onset is defined as the beginning of prostate cancer in an individual. Note that the time of onset for an individual cannot be directly determined.



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)



University of Michigan
define TURP

DEFINE TURP

Transurethral resection of the prostate (TURP) is a surgical procedure performed to treat benign prostate hyperplasia (BPH) and urinary obstruction symptoms.



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)



University of Michigan
Introduce Within Stage Shift

INTRODUCE WITHIN STAGE SHIFT

See [Delay_Time_Approach](#) for modeling of within-stage shift.



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)

AGE AT TUMOR ONSET

Age at tumor onset

We use a Weibull distribution for the baseline age at tumor onset. Its baseline hazard function is given by

$$h_O(y) = s_O \left(\frac{\Gamma(1 + 1/s_O)}{\mu_O} \right)^{s_O} y^{s_O-1},$$

where y is the age past 50. In the above expression Weibull distribution is parameterized through the mean μ_O and the shape parameter s_O related to the coefficient of variation

$$\sqrt{\frac{\Gamma(1 + \frac{2}{s_O})}{\Gamma^2(1 + \frac{1}{s_O})} - 1}.$$

Included in the model is a trend function $T_O(t)$ that depends on calendar time. This function exerts a multiplicative effect on the baseline hazard so that the hazard of tumor onset depends on age and birth cohort

$$\lambda_O(y|x) = h_O(y)T_O(x + y).$$

The trend is used to model possible changes in the pattern of the disease onset with calendar time due to unspecified factors such as changes in diet, environment and biology of the disease. Note that it is hardly possible to give a biological definition for the tumor onset. From the modelling prospective, tumor onset represents the earliest point in time where cancer could be detected by screening. For this reason changes in detection technology, practice of biopsies for the disease following a positive screens and other diagnostics management issues may also affect the definition. Changes in such practices that are not modelled in a mechanistic fashion are thought of as part of the trend function. We used truncated linear trend functions in data analysis.

(This is extracted from an early draft of¹.)

REFERENCES:

- ¹ Tsodikov, A., Szabo, A., Wegelin, J. "A population model of prostate cancer incidence" in *Statistics in Medicine* 2006; 25: 16: 2846-2866



Readers Guide
Model Overview
Assumption Overview
Parameter Overview
Component Overview
Output Overview
Results Overview
Key References



University of Michigan
define Sojourn Time

DEFINE SOJOURN TIME

Sojourn time is defined as the potential (other risks removed) time from tumor [onset](#) to its clinical diagnosis. Thus it is the duration of the preclinical stage in the absence of screening. We speak of the sojourn time distribution even for individuals who receive screenings. In this way we model the competing risks of clinical and screening diagnosis.



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)



SOJOURN TIME DISTRIBUTION

The sojourn time distribution.

Sojourn time is defined as the potential (other risks removed) time from tumor onset to its clinical diagnosis. A Weibull distribution with mean μ_{CDx} and shape parameter s_{CDx} is used to model the baseline sojourn time hazard. Two effects can be imposed on the baseline sojourn time distribution:

- Age. Sojourn time may be affected by age for various reasons. Tumor growth biology may depend on the age of the person. Also, tumors developing at a younger age may represent a special subtype that can have different progression characteristics. To model age dependency, the mean sojourn time is regressed on the age at tumor onset y as $\mu_{CDx} \exp(-\beta_{CDx}y)$, where the parameter β_{CDx} models the correlation between the sojourn time and the onset time.
- Secular trend. Sojourn time may be affected by changes in the practice of cancer detection other than the studied modality of screening. Most notably, before PSA was introduced, prostate cancer was often detected as a result of surgery (Transurethral Resection of the Prostate, [TURP](#)) for benign prostate disorders¹. Other changes in prostate cancer awareness in the population and detection practices may have contributed to a trend of increasing incidence observed before PSA was introduced. These trends in calendar time are modelled using a multiplicative trend function $T_{CDx}(t)$ acting on the baseline sojourn time hazard.

We have the sojourn time hazard in the form

$$\lambda_{CDx}(\xi|x, y) = h_{CDx}(\xi|y)T_{CDx}(x + y + \xi),$$

where x is the birth year, y is age (past 50) at tumor onset, ξ is time since tumor onset, and $h_{CDx}(\xi|y)$ is Weibull hazard with shape parameter s_{CDx} and mean $\mu_{CDx} \exp(-\beta_{CDx}y)$.

REFERENCES:

- ¹ Merrill, R.M., Feuer, E. J., Warren, J. L. , Schussler, N. , Stephensons, R. A. "Role of transurethral resection of the prostate in population-based prostate cancer incidence rates" in American Journal of Epidemiology 1999; 150: 8: 848-860



University of Michigan
define Delay Time

DEFINE DELAY TIME

Delay time is defined as the duration of the [latent_disease_stage](#), i.e., the time from onset until detection of cancer by any means, including PSA screening or clinical detection.



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)



University of Michigan
define Latent Disease Stage

DEFINE LATENT DISEASE STAGE

The **latent disease stage** is defined as the time when an individual has cancer but the cancer has not yet been detected by any means.



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)



INCIDENCE MODEL

INCIDENCE MODEL

We use the classical three-stage model of the natural history of a chronic disease¹. Prostate cancer is a result of an irreversible transition of the disease through three consecutive stages: disease free stage, pre-clinical stage and clinical stage. The time spent in disease-free stage is characterized by the age Y (a random variable) at onset of the disease. In the pre-clinical stage disease is asymptomatic and can be detected by a screening test. The duration of the preclinical stage in the absence of screening (a random variable) is termed the sojourn time. If undetected by screening, the disease can either reach the clinical stage or, alternatively, the event of clinical diagnosis is precluded by a competing risk other than the disease of interest.

The distribution of any random duration can be specified by one of the following functions: a hazard function (h.f., λ or h), a survival function (s.f., G), a distribution function (d.f., \overline{G}), or a probability density function (p.d.f., f). Dependent on the situation, we will use the most convenient representation. Denote age by a , calendar year by t , year of birth by x , and time since tumor onset by ξ . We will follow the above notation unless noted otherwise. Prostate cancer incidence $\lambda_I(a, t)$ by age and year can be written as $\lambda_I(a|t - a)$ where $\lambda_I(a|x)$ is the h.f. for cancer diagnosis for the x -birth cohort. Clearly,

$$\lambda_I(a|x) = \frac{f_I(a|x)}{G_I(a|x)}.$$

The functions f_I and G_I are in fact represented by a fairly complex mixture model. It is clear that cancer incidence is a convolution of two generally dependent survival times: age at tumor onset Y and duration of the latent disease stage T .

$$f_I(a|x) = \int_0^a f_I(a - y|x, y) f_O(y|x) dy,$$

where $f_I(\xi|x, y)$ is a conditional p.d.f. of T , and f_O is the p.d.f. of Y . Generally, $f_I(\xi|x, y)$ is an average over random patterns of screening operating in the population. It is clear that T is a result of two dependent competing risks: the one associated with natural clinical diagnosis through symptoms and the one associated with detection through screening. Dependency between the two risks is a consequence of natural detection and screen-based detection risks sharing the same disease development process in the subject. This dependency is modelled through the concept of shared mixed effect (frailty) Hougaard², represented by Y . Conditional independence of potential risks of natural and screen-based detection, given Y gives

$$G_I(\xi|x, y) = G_{CDx}(\xi|x, y) G_{SDx}(\xi|x, y),$$

where G_{CDx} is the s.f. of time to clinical diagnosis (CDx) in the absence of screening (the sojourn time), and G_{SDx} is the s.f. of the potential time to screen-based diagnosis (SDx). Note, that G_{SDx} in our model corresponds to a continuous distribution as it is represented as a continuous mixture over random screening schedules in the



population. Since incidence of prostate cancer before the age of 50 is negligible, we will associate the birth year x with the year in which the man turns 50. Weibull distribution with mean μ_O and the shape parameter s_O is used for the baseline age at tumor onset. Weibull distribution with mean μ_{CDx} and shape parameter s_{CDx} is used to model the baseline sojourn time hazard h_{CDx} . Two effects are imposed on the baseline sojourn time distribution, age dependence and a secular trend. To model age dependency, the mean sojourn time is regressed on the age at tumor onset y as $\mu_{CDx} \exp(-\beta_{CDx}y)$, where the parameter β_{CDx} models correlation between the sojourn time and the onset time. Secular trend models alterations in the practice of cancer detection other than the studied modality of screening. Most notably, in the pre-PSA era, many prostate cancers were incidentally detected through TURP

³. Secular trend is introduced as a multiplicative effect

REFERENCES:

- ¹ Zelen, M., Feinleib, M. "On the theory of screening for chronic diseases" in Biometrika 1969; 56: : 601-614
- ² Hougaard, P. "Frailty models for survival data" in Lifetime Data Analysis 1996; 1: : 255-274
- ³ Merrill, R.M., Feuer, E. J., Warren, J. L. , Schussler, N. , Stephensons, R. A. "Role of transurethral resection of the prostate in population-based prostate cancer incidence rates" in American Journal of Epidemiology 1999; 150: 8: 848-860



PSASCREENING MODEL

The PSA screening model

The National Cancer Institute's Statistical Research and Applications Branch has developed a simulator for PSA schedules for arbitrary birth cohorts in the 1916--2000 box. This simulator uses data from the National Health Interview Survey (NHIS)¹ and Surveillance, Epidemiology and End Results (SEER) -- Medicare linked database². To extrapolate the data beyond the original age--year box, generalized additive models (R procedure `\em gam`) were used to smooth the data. A logistic regression model was used for smoothing with the additive main effects of age a and calendar year t represented by thin plate regression splines³. No interaction smooth terms were specified. Shown in [Figure_3b](#), below, is an estimate for the risks of first $\lambda_{1S}(a, t)$ and secondary $\lambda_{2S}(a, t)$ PSA tests.

It is clear from the figure that the risk of secondary PSA test is several times higher the one for the first test. This observation prompted the development of the two-stage model for screening based detection described in [Modeling Cancer Detection Through Screening-B](#) and⁴. Frequency of PSA testing by age increases initially as the man enters the risk zone for prostate cancer. However for the older ages a decreasing pattern is observed perhaps because of limited residual life expectancy and associated diminishing relevance of detection of prostate cancer. Dissemination by calendar year is different for the first and secondary tests. In men who have been screened at least once the frequency increases as PSA is introduced into practice and the surface settles at stable values in the nineties. The risk of getting the first test by calendar year shows a spike in early nineties and settles at a lower level later showing a decreasing pattern in the late nineties. This phenomenon deserves further study. The effect could be a consequence of heterogeneity in people's acceptance of PSA testing. The group of men showing compliance for PSA testing is dissipating with time as such men get tested and leave the set of men "at risk" for the first test. Another explanation might be that the recent decline in the frequency of new PSA tests is associated with a dissemination of knowledge of various controversial issues surrounding screening for and treatment of prostate cancer.

REFERENCES:

- ¹ National Center for Health Statistics "National Health Interview Survey (NHIS)." 2004;
- ² National Cancer Institute "Surveillance Epidemiology and End Results (SEER) - Medicare linked database" 2002;
- ³ Wood, S. "Thin plate regression splines" in Journal of the Royal Statistical Society Series B 2003; 65: : 95-114
- ⁴ Tsodikov, A., Szabo, A., Wegelin, J. "A population model of prostate cancer incidence" in Statistics in Medicine 2006; 25: 16: 2846-2866



SURVIVAL COMPONENT

SUMMARY

This document describes the survival component of the model.

OVERVIEW

The survival component is an analytic model that describes the relationship between a set of covariates and a survival curve (alternatively a hazard function, a distribution of time to failure, or a density).

DETAIL

The model yields differing survival curves depending on the following covariates.

- stage of cancer
- tumor grade
- calendar year of diagnosis
- age of patient
- therapy (the integration of this covariate into the model remains as future work.)

Here we present the main results. These results permit us to use composition to build flexible semiparametric survival models (nonlinear transformation models) and use them for estimation and hypothesis testing.

Nonlinear transformation models

Let $\gamma(x | \beta, z)$ be a parametrically specified distribution function with the x -domain of $[0, 1]$. Let $F(t)$ be a nonparametrically specified baseline survival function. A semiparametric regression survival model is called a Nonlinear Transformation Model if its survival function can be represented as a composition

$$G(t | \beta, z) = \gamma \{F(t) | \beta, z\} = (\gamma \circ F)(t | \beta, z).$$

The NTM class and associated estimation procedures were developed by Tsodikov¹. The key requirement that ensures monotonicity and convergence of the estimation algorithms (see Estimation Algorithm) is that of nondecreasing $\Theta(x | \cdot)$, where where $\gamma^{(c)}(x | \cdot) = \partial^c \gamma(x | \cdot) / \partial x^c$, $c = 0, 1, \dots$, $\gamma^{(0)}(x | \cdot) = \gamma(x | \cdot)$. Using frailty models analogy, $\Theta(F | \cdot, c)$ can be interpreted as a surrogate of the posterior risk for a subject observed with an event at time t , where $c=0$ if right censored, $c=1$ if failed.

Model building by composition

If γ_θ and γ_η are two different NT models with predictors θ , and η , respectively, then

is a new semiparametric model with two predictors θ and η . The fact that NTM--generating functions $\gamma(x | \cdot)$ are all defined on $x \in [0, 1]$ and have the range in the same interval allows us to compose as complex a hierarchical model as needed. Moreover, we proved that operation of composition preserves the key property of nondecreasing Θ observed in frailty models \citep{tsomodelbuilding}. We also derived a chain rule that allows us to specify Θ for the compound model based on Θ -functions of the submodels

As we will see in the next section, knowledge of Θ is all that is needed to specify an estimation procedure.

Estimation algorithm

Let $t_i, i = 1, \dots, n$ be a set of times, arranged in increasing order, $t_{n+1} := \infty$. Associated with each t_i is a set \mathcal{R}_i of subjects $j \in \mathcal{R}_i$ at risk, with covariates z_{ij} . For any function $A(t)$, let $A_i = A(t_i)$, $\Delta A_i = |A(t_i) - A(t_i - 0)|$. The following method (QEM) is used to obtain the profile likelihood.

where $\{F^{(k)}\}$ and $\{H^{(k)}\}$ are sequences of functions generated by the self-consistency equation (Equation QEM), D_m is the number of failures at t_m and β is a vector of regression coefficients.

It can be shown that if Θ is nondecreasing, each update of H using the self-consistency equation (Equation QEM) strictly improves the likelihood, given β . This guarantees convergence of the sequence of likelihood values $\ell\{\beta, H^{(k)}\}$ to the profile likelihood of β , and of the sequence $\{H^{(k)}\}$ to H^* , the fixed point of $(\backslash \text{ref}\{\text{qem}\})$, under fairly general conditions.

Under a frailty model, the procedure (Equation QEM) is an EM algorithm based on imputation of a missing predictor by its conditional expectation, given observed data, represented by $\Theta(F | \beta, z, c)$. Under an NT model, the procedure works as a Quasi-EM algorithm without the missing-data interpretation.

Profile information matrix

To obtain confidence intervals and tests of statistical hypotheses for regression coefficients, we developed a solution for the exact observed profile information matrix of β^2 . As the number of parameters of a semiparametric model is potentially unlimited, obtaining the inverse of the full information matrix becomes computationally prohibitive, and a profile information matrix would be very useful. The profile information matrix can be expressed as

where $H = (\Delta H_1, \dots, \Delta H_n)^T$ and

$$I_{ab} = - \left. \frac{\partial^2 \ell(\beta, H)}{\partial a \partial b} \right|_{(\beta, H^*(\beta))}$$

for any two vectors a and b , where ℓ is a log-likelihood and H^* is the fixed point of the self-consistency equation. Notice that $I_{\beta\beta}^*$ has dimension $d \times d$ with $d = \dim(\beta)$, therefore only a small matrix needs to be inverted in order to get an estimator of the covariance matrix of regression coefficients. The downside of $(\backslash \text{ref}\{\text{ipr1}\})$ is that since $H^*(\beta)$ is defined implicitly, so is the potentially large Jacobian matrix $\partial H^* / \partial \beta$. Therefore, the Jacobian is generally unavailable in a closed form. In the NTM case the problem reduces to solving a system of linear equations $(D + R)x = b$, where x represents a column--vector of the Jacobian, D is an $n \times n$ diagonal matrix with



diagonal elements $d_i \neq 0, i = 1, \dots, n$, $R = (R_{kl})$ is a $n \times n$ matrix, $R_{kl} = \sum_{i=\max\{k,l\}}^n a_i a_i$, $i = 1, \dots, n$ are real numbers, and b be an n -dimensional vector. The main result used to obtain $\partial H^*/\partial \beta$ is as follows. Let the functions $\varphi_k : \rightarrow, k = 1, \dots, n$ be defined recursively as $\varphi_n(y) = b_n/d_n - a_n/d_n y$, $\varphi_k(y) = \left(b_k - \sum_{i=k}^n a_i y + \sum_{l=k+1}^n \sum_{i=k}^{l-1} a_i \varphi_l(y) \right) / d_k$, $k = n-1, \dots, 1$. Now, let $\tilde{\varphi}$ be the function given by $\tilde{\varphi}(y) = \sum_{k=1}^n \varphi_k(y)$. The solution to the system of equations $(D + R)x = b$ is the n -dimensional vector $x = (\varphi_1(\tilde{y}), \dots, \varphi_n(\tilde{y}))$, where $\tilde{y} = \tilde{\varphi}(0)/(1 + \tilde{\varphi}(0) - \tilde{\varphi}(1))$.

RELEVANT ASSUMPTIONS

See [Assumption Overview](#).

RELEVANT PARAMETERS

Recall that the term "parameter," in the language of this model profile environment, actually refers to a model input.

The inputs to the survival model consist of data from the SEER (Surveillance, Epidemiology and End Results) database, which includes approximately 350,000 men diagnosed with cancer³. More information on these data may be found by clicking the Details link in the footnote.

Each man's covariates (age, stage, grade, etc.) enter individually into the survival model. In this, this survival model differs from the incidence model (see [Incidence Component](#)).

RELEVANT COMPONENTS

The components of the survival model are results that allow us to build flexible semiparametric survival models and use them for estimation and hypothesis testing. Many of these results are new discoveries, developed under this project. Details may be seen above, on this page.

DEPENDENT OUTPUTS

[Mortality Component](#)

RELEVANT RESULTS

Methods. Advances were made in statistical methods in the development of this component of the model. For more information, see the Relevant Components section of this document.

Parameter estimates and prediction. The output of the Survival Component is combined with the output of [Incidence Component](#) and used as input to the [Mortality Component](#), leading to predicted mortality and an estimate of the difference in mortality that is due to the introduction of PSA testing. See [results Age Adjusted Incidence Mortality](#) and [Analysis Of Population Data](#).

REFERENCES:

¹ Tsodikov, Alex "Semiparametric models: a generalized self-consistency approach" in Journal of the Royal Statistical Society, Series B 2003; 65: : 759–774



University of Michigan
Survival Component
References:

- ² Tsodikov, A., Garibotti, G. "Profile information matrix for nonlinear transformation models" in Lifetime Data Analysis 2007; 13: 1: 139-159
- ³ National Cancer Institute "Surveillance Epidemiology and End Results (SEER) - Medicare linked database" 2002;



MORTALITY MODEL

THE MORTALITY MODEL

Let z be the vector of clinical covariates observed at diagnosis. The results of [Survival Component](#) allow us to study different survival models $G_M(\tau|z, a, t)$ (s.f.), where τ is the survival time post-diagnosis, a is age at diagnosis, t is calendar year of diagnosis, and z represents clinical covariates. We found that the proportional hazards model does not provide a good fit for the data by stage and grade. The list of adequate survival models for prostate cancer includes the PHPH cure model^{1,2} and the proportional odds (PO) model⁴. We prefer the PO model with one linear predictor over the PHPH cure model with two predictors by the AIC model selection criterion. A test for the PO assumption vs. the PH assumption using the Gamma frailty model with covariates in shape and scale parameters of the frailty distribution can be found in Tsodikov2004. Confidence intervals for odds ratios of stage can be found in⁵.

In [Incidence Model](#) we presented a marginal model for cancer incidence by age a and year of diagnosis t . The marginal p.d.f. $f_I(a|x)$ for the x -cohort, $x = t - a$ resulting from this model can be partitioned into z -specific fractions using a regression of z on age and year, $f_z(z|a, t)$, so that $f_I(a, z|t - a) = f_I(a|t - a)f_z(z|a, t)$.

In practice, to specify f_z we use a categorical effect of t and define z as a categorical prognostic variable based on stage and grade. Cutpoints on PSA value and its velocity at diagnosis can be used to extend z . The period categorical variable associated with t serves as a surrogate of PSA utilization affecting z -shift and within--stage shift of survival with the introduction of PSA. We consider this approach preliminary in that it only pertains to the actually observed utilization pattern and does not easily generalize to hypothetical PSA impact scenarios (unlike the incidence model). As part of model improvement in this competing continuation application we plan to provide a more sophisticated link between the incidence and survival models that is necessary to address the specific aims of this project.

Now, mortality in year t_M at the age of a_M , $\lambda_M(a_M, t_M) = \lambda_M(\tau|x)$, where $\tau = t_M - t = a_M - a$ and $x = t - a = t_M - a_M$, is derived from the prostate cancer specific survival function counted out from birth, represented by the following convolution:

REFERENCES:

- ¹ Tsodikov, A. "Semiparametric models of long- and short-term survival: An application to the analysis of breast cancer survival in utah by age and stage" in *Statistics in Medicine* 2002; 21: : 895-920
- ² Tsodikov, A., Ibrahim, J.G., Yakovlev, A.Y. "Estimating cure rates from survival data: An alternative to two-component mixture models" in *Journal of the American Statistical Association* 2003; 98: : 1063-1078
- ³ Tsodikov, Alex "Semiparametric models: a generalized self-consistency approach" in *Journal of the Royal Statistical Society, Series B* 2003; 65: : 759-774
- ⁴ Tsodikov, A. "Using composition to build semiparametric survival models" in *Statistical Modelling* 2006; Submitted:
- ⁵ Tsodikov, A., Garibotti, G. "Profile information matrix for nonlinear transformation models" in *Lifetime Data Analysis* 2007; 13: 1: 139-159



University of Michigan
Mortality Model
References:



University of Michigan
base Case PSA



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)

BASE CASE PSA

The base case PSA simulation program can be used to simulate life histories of the times that individual men undergo PSA tests. The simulator is based on data from the National Health Interview Survey¹ and the Surveillance, Epidemiology and End Results (SEER) -linked database². The simulation is based on two submodels for the "risk" of PSA test, both of which depend on age and calendar year. The first submodel is a survival model for the time to first PSA test among men who have not yet had a test. The second submodel is a non-homogeneous Poisson process model for the schedule of subsequent PSA tests in men who have already had at least one PSA test. For further information, see [figure First PSAtest Secondary PSAtest](#).

REFERENCES:

- ¹ National Center for Health Statistics "National Health Interview Survey (NHIS)." 2004;
- ² National Cancer Institute "Surveillance Epidemiology and End Results (SEER) - Medicare linked database" 2002;



University of Michigan
Seer_Medicare



Readers Guide
Model Overview
Assumption Overview
Parameter Overview
Component Overview
Output Overview
Results Overview
Key References

SEER_MEDICARE

REFERENCE

National Cancer Institute (2002), "Surveillance Epidemiology and End Results (SEER) - Medicare linked database"

URL:

<http://healthservices.cancer.gov/seermedicare/>

NOTES AND DISCUSSION

The SEER data used in the current study consist of two parts:

- Population data: an age-by-year table with a count in each cell of the number of men at risk for prostate cancer
- Survival data: one row for each patient, with several variables such as age at diagnosis, stage, grade, and outcome

The data are available only from the following locations: San Francisco-Oakland, Connecticut, Detroit, Hawaii, Iowa, New Mexico, Seattle, Utah, and Atlanta. They include all prostate cancer cases diagnosed in these regions, approximately 350,000 men.

CategoryReferences



LIKELIHOOD IN THE INCIDENCE MODEL

LIKELIHOOD IN THE INCIDENCE MODEL

Observed data for the incidence model include a count $P(a, t)$ of people at risk of cancer and a count of cancer cases $C(a, t)$ by age and year. The conditional likelihood of the data is built as a product of conditional probabilities of cancer detection given that the subject is in the risk set for each a, t combination from the observed box. Except for terms that do not depend on the model parameters, the likelihood takes the form

$$\ell = \sum_{a,t} C(a, t) \log \lambda_I(a, t) - P(a, t) \lambda_I(a, t).$$

Note that the same likelihood would result if we assumed that C is Poisson distributed with expectation $P\lambda_I$ and that $C(a, t)$ represent independent random variables for different (a, t) pairs (which is not the case in Equation Likelihood). Maximum likelihood inference is used to obtain point estimates and confidence intervals for the model parameters entering λ_I . Maximization of the likelihood can be regarded as minimizing a certain distance between the empirical incidence C/P and its model-based counterpart λ_I .



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)



University of Michigan
Human Mortality Database

HUMAN MORTALITY DATABASE

REFERENCE

Wilmoth, John R. (Director), Shkolnikov, Vladimir Shkolnikov (Co-Director), (2003), "Human Mortality Database (HMD)."

URL:

<http://www.mortality.org/>

NOTES AND DISCUSSION

CategoryReferences



Readers Guide
Model Overview
Assumption Overview
Parameter Overview
Component Overview
Output Overview
Results Overview
Key References



INCIDENCE COMPONENT

SUMMARY

This document summarizes the incidence component of the model.

OVERVIEW

The incidence component is an analytic model for the time until prostate cancer diagnosis of a randomly-selected man from the population at risk. This model is expressed in terms familiar from survival analysis, and thus may be expressed as a hazard function, a density, a cumulative distribution function, or a survival function. A likelihood-based estimation procedure is part of this component.

DETAIL

The incidence model uses population data to estimate [parameters](#) that characterize prostate cancer incidence as a function of age and calendar year.

RELEVANT ASSUMPTIONS

Please see [Assumption Overview](#).

RELEVANT PARAMETERS

Recall that the term "parameter," in the language of this model profile environment, actually refers to a model input. The inputs to the incidence model consist of data and of [parameters](#) (in the statistical sense of the word) belonging to a model that has been specified and estimated independently.

The incidence model currently uses data from nine areas of the United States: San Francisco-Oakland, Connecticut, Detroit, Hawaii, Iowa, New Mexico, Seattle, Utah, and Atlanta. Before being used in the estimation procedure, the data are summarized by age (a) and calendar year (t).

The particular variables summarized by age and calendar year, and used as input to the model, are:

- The number of men at risk for prostate cancer, $P(a, t)$ for each age a and calendar year t .
- The number of new prostate cancer cases, $C(a, t)$ by age a and calendar year t , obtained by computing summaries from the SEER (Surveillance, Epidemiology and End Results) database¹.

In addition, the estimation procedure uses a prior estimate of the distribution of PSA utilization in the population. This distribution is based on an algorithm that can be used to simulate life histories of the times that individual men undergo PSA tests. For more information see [base Case PSA](#).

RELEVANT COMPONENTS

- Distribution of time until tumor [onset](#) (see [Incidence Model](#))
- Duration of the latent disease stage (also called delay time), which is further broken down into the following competing risks:



- Risk of clinical diagnosis (see [Incidence Model](#))
- Risk of screen-based diagnosis (see [Modeling Cancer Detection Through Screening](#))

(Other means of detection, such as [TURP](#), are not currently included in the model.)

- Likelihood and estimation algorithm (see [Likelihood In The Incidence Model](#))

DEPENDENT OUTPUTS

[Mortality Component](#)

RELEVANT RESULTS

- Predictions of incidence with and without PSA testing (see [results Age Adjusted Incidence Mortality](#)), and consequently
- Estimation of the difference in incidence that is due to PSA testing.

REFERENCES:

- ¹ National Cancer Institute "Surveillance Epidemiology and End Results (SEER) - Medicare linked database" 2002;



University of Michigan
Mortality Component



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)

MORTALITY COMPONENT

SUMMARY

This document describes the mortality component of the model.

OVERVIEW

The [Mortality Component](#) yields estimates of mortality by calendar year, age, and presence or absence of PSA testing.

DETAIL

See [Mortality Model](#).

RELEVANT ASSUMPTIONS

See [Assumption Overview](#).

RELEVANT PARAMETERS

Recall that the term "parameter," in the language of this model profile environment, actually refers to a model input, not a [parameter](#) in the classical statistical sense.

The inputs to this component are the outputs of the [Incidence Component](#) and [Survival Component](#).

This component does not use population mortality data as input.

RELEVANT COMPONENTS

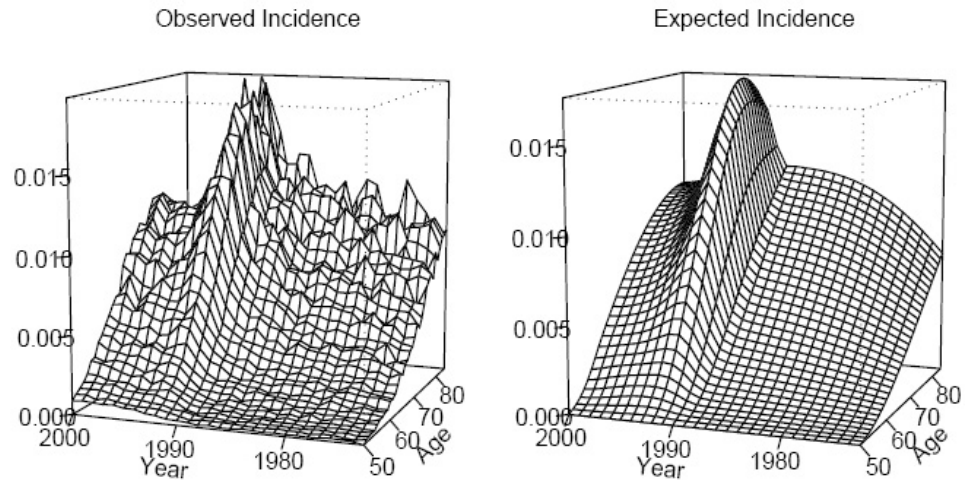
DEPENDENT OUTPUTS

RELEVANT RESULTS

- Predictions of prostate cancer-specific mortality with and without PSA testing (see [results Age Adjusted Incidence Mortality](#)) and consequently
- Estimation of the difference in mortality that is due to PSA testing.



INCIDENCE FIGURE



Incidence Figure. Prostate cancer incidence (rate per person). Observed incidence of prostate cancer is displayed on the left. This is a histogram empirical estimate obtained by dividing incident cancer cases by the population at risk, for each age and each calendar year. Expected incidence is displayed on the right, also by age and calendar year, as predicted by the model. The model captures the basic pattern of prostate cancer incidence. The spike occurring with the introduction of PSA testing gets more pronounced with increasing age, except in very old men. The decrease in older men is a consequence of the fact that latent prevalence accumulates with age.



University of Michigan
define Lead Time



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)

DEFINE LEAD TIME

Lead time refers to the amount by which detection of prostate cancer is advanced due to PSA screening. It adds to the observed survival time even if early detection and treatment were of no benefit. The lead-time effect targets patients who would still be detected later without screening. This effect could result in apparently improved short-term survival even if there were no mortality benefit.



RESULTS OVERDIAGNOSIS LEAD TIME

Estimates of lead time and overdiagnosis reveal the potential natural history of the disease and of population screening exposure over the lifespan of an individual. These parameters are estimated by birth cohort, and presented in [Figure_5](#). Overdiagnosis can be variously defined as a fraction of all detected cancers (the solid curve in the left panel of [Figure_5](#)), or as a fraction merely of screen-detected cancers (the dashed curve).

Recall that the horizontal axis represents the year of a man's fiftieth birthday, so that older men are represented toward the left of each panel in [Figure_5](#) and younger men toward the right. In younger cohorts, more of the cohort life span falls in the PSA era. This leads to a pattern of increasing lead time and increasing overdiagnosis among all detected cancer patients as we move toward the right in each panel (solid curves). For men entering the age risk zone for prostate cancer at the present time, the model predicts about a six-year mean lead time and 25% overdiagnosis among all detected patients.

Overdiagnosis in screen-detected cases is represented by the dashed curve in the left panel. This must always be higher than the solid curve, because screen-detected cases are a proper subset of all cancer cases and thus there is a smaller denominator in computing the fraction. But in addition to being higher than the solid curve, the dashed curve reveals a trend in the opposite direction. This can be understood as follows. Men whose fiftieth birthday occurs in the 1950s were already very old in the PSA era. A prostate cancer detected by screening in a man of this age has a high probability of being overdiagnosed, because of his very small expected residual lifetime. In men who are younger during the PSA era, on the other hand, the pool of screen-detected cases include includes many cancers that would have surfaced clinically in the man's residual lifetime if the man had not received a PSA test. These relevant cancers reduce the proportion of overdiagnosed cancers in younger men. Overdiagnosis in screen-detected cases settles at about 30% for men who reach their fiftieth birthday during the PSA era.



University of Michigan
define Overdiagnosis



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)

DEFINE OVERDIAGNOSIS

Overdiagnosis. A large proportion of prostate cancers identified through screening would never be detected in the absence of screening. This phenomenon is called overdiagnosis. Screening brings such cancers to the surface predominantly in the localized stage of the disease, leading to an apparent "favorable" stage shift. Overdiagnosis has multiple consequences. It leads to over-treatment of men who would never be detected without screening. Also, it modifies apparent estimates of post-treatment survival as over-diagnosed cases appear to be "cured." Injection of overdiagnosed cases into the pool of all prostate cancer presentations at diagnosis changes the distribution and the meaning of clinical covariates in men diagnosed with prostate cancer in the PSA era. Overdiagnosis could lead to apparently improved long-term survival of patients with localized stage of the disease even if there were no mortality benefit.



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)

MEAN POSTERIOR DELAY TIME MARGINAL INCIDENCE MODEL

[Figure_8](#) shows an estimate of delay time computed from SEER data ([Analysis Of Population Data](#)). Introduction of PSA testing is associated with earlier detection, and the older the man the more so. The slight decrease in delay time in the no-PSA prediction is a transient process resulting from freezing the pre-PSA trend estimates in the year 1988.

For more information on the model from which this estimate was obtained, see [Incidence Model](#).

For more information on delay time and its integration into the model, see [Delay_Time_Approach](#).



WITHIN STAGE SHIFT BY DELAY TIME

[Figure_9](#) shows that the delay-time approach, developed in the current project, captures within-stage-and-grade shift. It should be stressed that delay time was computed without using survival data, and represents an independent significant prognostic factor for post-treatment survival, particularly in the localized-regional. These results can be compared with [Figure_2](#), where a similar effect is expressed by year of diagnosis. For estimates of delay time as a function of age and calendar year, see [Mean Posterior Delay Time Marginal Incidence Model](#). For more information on the model from which these estimates were obtained, see [Delay_Time_Approach](#).



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)

RESULTS AGE ADJUSTED INCIDENCE MORTALITY

In [Figure_6](#), model predictions are displayed for age-adjusted incidence and mortality, along with their empirical estimates. To generate predictions in the counterfactual case of no PSA testing, all trend functions were frozen at a constant in the year 1988, and PSA sensitivity was set at zero. The mortality figure (right) indicates that the introduction of the PSA test has led to a decline in mortality.

For a description of the data analysis that yielded these results, see [Analysis Of Population Data](#).

For a deeper understanding of the model components on which the results are based, see [Incidence Model](#) and [Mortality Model](#).



ANALYSIS OF POPULATION DATA

ANALYSIS OF POPULATION DATA

The SEER database was used to obtain data on more than 350,000 cases of prostate cancer diagnosed in nine areas of the United States (San Francisco-Oakland, Connecticut, Detroit, Hawaii, Iowa, New Mexico, Seattle, Utah, Atlanta) as well as population count files corresponding to those cases. We use the modeling box corresponding to age interval [50,85] and calendar year interval [1973-2000]. Age distribution in the U.S. population in year 2000 for men over 50 is used as a standard when age-adjusted characteristics are reported. Risk of death from other causes (used in estimates of lead-time and overdiagnosis) was derived from the [Human Mortality Database](#)¹.

As shown in [Figure_1](#), incidence of prostate cancer before the introduction of PSA showed an increasing trend in calendar time reportedly related to TURPs². In order to model this effect, a linear trend was specified for the sojourn time model ([Equation LCDX](#)) for the period 1973-1987, saturating in 1988. The parameter c specifies the slope of the trend during 1973-1987. We did not have a compelling evidence for changes in the onset time distribution over time, and $T_O(t)$ was removed from the model. Also, we did not find any improvement in the fit from introducing a correlation between age at onset and the sojourn time, and this term was removed from the model. PSA sensitivity was specified as an increasing function of the time since tumor onset. When fitting the model, the estimate settled at 100% sensitivity. Likelihood was maximized by the Powell's method (Himmelblau1972) of conjugate directions. Confidence intervals for the model parameters are based on Likelihood Ratio and inverting the profile likelihood surface for each parameter. Estimates of key model parameters and the corresponding confidence intervals are shown in [Table_1](#).

Note that the estimated mean age at tumor onset goes well beyond the normal human lifetime. This is a consequence of the fact that only a proportion of men would ever develop prostate cancer in their life span. Shown in [Figure_4](#) is a histogram empirical estimate of prostate cancer incidence $C(a, t)/P(a, t)$ and its model--predicted counterpart $\lambda_I(a, t)$ by age and calendar year.

The model captures the basic pattern of prostate cancer incidence. The spike effect in the incidence occurring with the introduction of PSA gets more pronounced with age except for very old people. This is a consequence of latent prevalence of the disease accumulating with age. Shown in [Figure_5](#) is an estimate of lead time and overdiagnosis. Both notions relate to the potential natural history of the disease and population screening exposure over the life span of an individual. Therefore we represent them by birth cohort. Overdiagnosis can be measured as a fraction relative to all detected cancers or to screen-detected cancers only. As we move the year of birth to the right, more and more of the cohort life span falls on the PSA era. This leads to an increasing pattern of lead time and overdiagnosis among all detected cancer patients (solid curves). For men entering the age risk zone for prostate cancer at the present



time, the model predicts about 6-year mean lead time and 25% overdiagnosis among all detected patients. Interestingly, overdiagnosis in screen-detected cases is a decreasing function of the birth year and settles at about 30% for the present era. Initially for a person born in the fifties only older ages are affected by PSA utilization. If detected at such an age, the case is very likely to be overdiagnosed. Indeed, if screening were ignored the disease would have little chance to surface because of the very small expected residual lifetime in older people. This is why the dashed curve in [Figure_5](#) (left) starts high. As we move the potential life history more and more under the PSA exposure, the pool of screen-detected cases gets enriched with relevant cancers that have advanced diagnosis due to PSA yet would surface clinically in their potential residual lifetime if PSA were not applied. Since screen-detected cases represent a subset of all cancer cases, overdiagnosis relative to screen-detected cases (the dashed curve) is always higher than the one relative to all cancer cases (the solid curve). Shown in [Figure_6](#) are model predictions for age-adjusted incidence and mortality and their empirical estimates. To generate predictions without PSA, all trend functions were frozen at a constant in year 1988, and PSA sensitivity was set at zero. The mortality figure (right) indicates that introduction of PSA test has led to mortality decline. Explaining this effect and its partitioning into fractions attributable to early detection and treatment is one of the emphases of this project.

The model was implemented in a prototype software package for Windows that brings incidence, survival, mortality and other model blocks into a common GUI shell that uses unified data input, output, menu and graphics structure. Shown in [Figure_7](#) are screen shots of the software.

REFERENCES:

- ¹ Wilmoth, John R. (Director), Shkolnikov, Vladimir Shkolnikov (Co-Director), "Human Mortality Database (HMD)." 2003;
- ² Merrill, R.M., Feuer, E. J., Warren, J. L. , Schussler, N. , Stephenson, R. A. "Role of transurethral resection of the prostate in population-based prostate cancer incidence rates" in American Journal of Epidemiology 1999; 150: 8: 848-860



TABLE_1

Parameter	Legend	Point Estimate	95% CI
μ_{CDx}	Mean baseline sojourn time	18.558	(18.345, 18.775)
s_{CDx}	Shape sojourn time	1.541	(1.5191, 1.5644)
c	Slope of trend for sojourn time	0.09354	(0.09068, 0.09641)
μ_O	Mean age past 50 at tumor onset	72.732	(72.498, 72.965)
s_O	Shape of age past 50 at tumor onset	1.6153	(1.6067, 1.6239)

[Table_1](#). Estimates of model parameters and confidence intervals. Time and age are measured in years.



MODELING CANCER DETECTION THROUGH SCREENING

MODELING CANCER DETECTION THROUGH SCREENING

Let $G_{SDx}(\xi|x, y)$ be the survival function of the potential time to screen-based diagnosis (SDx). This section is devoted to modeling this distribution.

For an arbitrary individual from the target population, consider the "risk" of getting his first screen. Age at first screen may be regarded as a survival time with the instantaneous risk represented by the hazard function $\lambda_{1S}(a, t)$ that depends on age and calendar year. An empirical histogram estimate for $\lambda_{1S}(a, t)$ can be obtained by dividing the number of subjects at the age of a receiving their first screen in year t by the total number of person-years with no evidence of the disease in the (a, t) cell. Of course, this estimate is inconsistent unless the data are grouped¹.

The probability of no screens by the age of a , G_{1S} , is a survival function obtained by integrating the hazard λ_{1S} over a life line on the so-called Lexis diagram²:

Denote by $\lambda_{2S}(a, t)$ the intensity of screening in subjects who have already had their first screen. The fact that the subject has had his first PSA test may identify him as a member of the group that enjoys a higher screening utilization for various reasons. Therefore, λ_{2S} is larger than λ_{1S} , as we see in [Figure_3](#).

Consider the unconditional probability $G_{2SDx}(\tau|x, a, y)$ that a subject is not diagnosed by screening in the age interval $[a, a + \tau]$, $a \geq y$. Under the assumptions stated in our Assumptions section, we have

$$G_{2SDx}(\tau|x, ay) = \exp \left\{ - \int_{\max(y-a, 0)}^{\tau} \lambda_{2S}(a + \zeta, x + a + \zeta) \alpha(\zeta + a - y) d\zeta \right\},$$

where $\int_a^b = 0$ for any $b \leq a$. The conditional probability of no screening diagnosis by the age of $y + \xi$, $G_{SDx}(\xi|x, y)$, takes the form

$$G_{SDx}(\xi|x, y) = G_{1S}(y + \xi|x) + \overline{G}_{1S}(y|x) G_{2SDx}(\xi|x, y, y) + \int_0^{\xi} \overline{\alpha}(\nu) f_{1S}(y + \nu|x) G_{2SDx}(\xi - \nu|x, y + \nu, y) d\nu.$$

The first term in this equation addresses the possibility of no screens by the age of $y + \xi$. The second term addresses the situation when the first screen occurs before onset of the disease at the age of y and no diagnosis is achieved through secondary screens that might happen in the age interval $(y, y + \xi)$. The third term accumulates the probability that cancer is missed at the first and secondary screens occurring after disease onset.



REFERENCES:

- ¹ Burr, D. "On inconsistency of Breslow's estimator as an estimator of the hazard rate in the cox model" in Biometrics 1994; 50: 4: 1142-1145
- ² Breslow, N. E., Lubin, J. H. , Marek, P., Langholz, B. "Multiplicative models and cohort analysis" in Journal of the American Statistical Association 1983; 78: 381: 1-12



University of Michigan
Publications



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)

PUBLICATIONS

- Broet, P., , Tsodikov, A., , De Rycke, Y., , Moreau, T.** (2004) Two-sample statistics for testing the equality of survival functions against improper semi-parametric accelerated failure time alternatives: An application to the analysis of a clinical trial in *Lifetime Data Analysis* 10 : , p 103-120
- Tsodikov, A., Garibotti, G.** (2007) Profile information matrix for nonlinear transformation models in *Lifetime Data Analysis* 13 : 1 , p 139-159
- Tsodikov, A., Ibrahim, J.G., Yakovlev, A.Y.** (2003) Estimating cure rates from survival data: An alternative to two-component mixture models in *Journal of the American Statistical Association* 98 : , p 1063–1078
- Tsodikov, A.** (2006) Using composition to build semiparametric survival models in *Statistical Modelling* Submitted :
- Tsodikov, A., Szabo, A., Wegelin, J.** (2006) A population model of prostate cancer incidence in *Statistics in Medicine* 25 : 16 , p 2846-2866
- Tsodikov, Alex** (2003) Semiparametric models: a generalized self-consistency approach in *Journal of the Royal Statistical Society, Series B* 65 : , p 759–774



University of Michigan
Over Diagnosis



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)

OVER DIAGNOSIS

A large proportion of prostate cancers identified through screening would never be detected in the absence of screening. This phenomenon is called overdiagnosis. Screening brings such cancers to the surface predominantly in the localized stage of the disease, leading to an apparent "favorable" stage shift. Overdiagnosis has multiple consequences. It leads to over-treatment of men who would never be detected without screening. Also, it modifies apparent estimates of post-treatment survival as over-diagnosed cases appear to be "cured." Injection of overdiagnosed cases into the pool of all prostate cancer presentations at diagnosis changes the distribution and the meaning of clinical covariates in men diagnosed with prostate cancer in the PSA era. Overdiagnosis would lead to apparently improved *long-term* survival of patients with localized stage of the disease even if there were no mortality benefit.



University of Michigan
Length Bias



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)

LENGTH BIAS

It has long been recognized that screening preferentially detects slower growing tumors¹. Slower growing tumors are likely to be associated with better prognosis. Among other consequences, length-bias effect would lead to apparently worsened survival of patients in distant stage under screening as compared to the unscreened population. Indeed, the pool of advanced tumors detected in the unscreened population is heterogeneous in terms of growth rates. With the introduction of screening some of the would-be distant cases will be detected earlier in a localized stage. These are likely to be the "best" slower growing fraction of the would-be distant cases. As a result, cases missed by screening that are still detected with distant disease under screening, show worse prognosis as compared to the unscreened population.

REFERENCES:

- ¹ Zelen, M., Feinleib, M. "On the theory of screening for chronic diseases" in Biometrika 1969; 56: : 601-614



University of Michigan
Lead Time

LEAD TIME

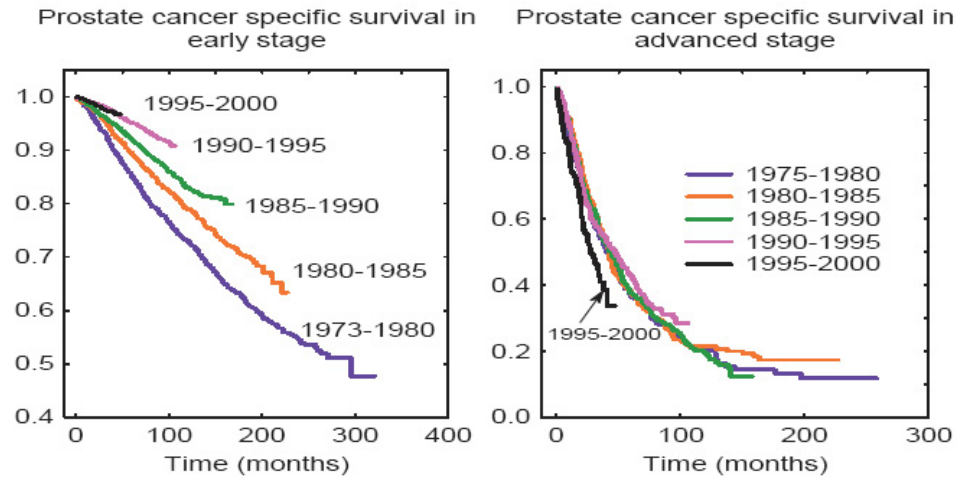
Lead-time measures an advance in the diagnosis of prostate cancer due to screening. It adds to the observed survival time even if early detection and treatment were of no benefit. The lead-time effect pertains to patients who would still be detected later without screening. Lead-time would lead to apparently improved *short-term* survival even if there were no mortality benefit.



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)



FIGURE_2



Figure_2. Within-stage shift. Prostate cancer specific survival by year of diagnosis and stage. The trend reflects improvement of prognosis in localized disease and worsening of prognosis in distant disease with dissemination of screening. Lead-time, length-bias and overdiagnosis provide part of the explanation for the within-stage shift. Data from the Surveillance, Epidemiology and End Results (SEER) database, National Cancer Institute.



DELAY_TIME_APPROACH

LINKING INCIDENCE AND SURVIVAL: THE DELAY TIME APPROACH

on the level of the subject, age at cancer diagnosis and survival post-diagnosis as well as cancer-specific lifetime are confounded by screening schedules. this confounding is expressed through lead-time, length bias and stage- and within-stage shift (see [Modeling Cancer Detection Through Screening](#)). As a consequence, as we discussed earlier, subjects with different screening schedules will have different distributions of age at diagnosis, and, clinical covariates being equal, they will still show different survival (within-stage shift). We have shown preliminary evidence that within-stage shift is a very significant effect in prostate cancer (Figure_2). The within-stage shift effect is a consequence of heretogeneity in the latent natural history of the disease and its strong effect on cancer detection processes. As a result, conditioning on different screening histories, other things equal, selects different subsets of natural histories of the disease. It then comes at no surprise that different natural histories are associated with different prognosis. In population data, person-level screening schedules are typically unavailable. Latent heterogeneity in the population data is much higher due to the contribution of uncertainty in screening schedules. In our population model, discussed in the Model Description above, the effect of screening on survival was modeled through the observed stage-shift and the within-stage shift adjusted for empirically through the categorized year of diagnosis variable. This allowed us to make mortality prediction within the observed period of 1973-2000, including a no-PSA predictive run performed by freezing all trend functions in 1988 and removing PSA from the incidence model. In this project we plan to improve the predictive potential of the model by linking population characteristics of interventions such as utilization of PSA, TURPs, Treatment, etc., to survival through natural history surrogates without using empirical variables such as year of diagnosis. This would make predictions a function of utilization characteristics and enable long-term predictions, optimization of interventions, unbiased assessment of treatment effects from population data, and many other model applications discussed below.

Our approach to linking incidence and survival using population data will be based on the concept of frailty. Conditional on the information available at diagnosis

1. Age at diagnosis, a , year of diagnosis, t ,
2. Clinical covariates observed at diagnosis, z ,
3. Dissemination of interventions over calendar time (rates of PSA testing, TURPs), $d(t)$,

we will derive the posterior distribution of the age at tumor onset Y . With the marginal incidence model discussed in [Incidence Model](#) (we continue using the notation introduced in this section), we have

$$Y \mid \{a, t, d(t)\} \propto f_O(y|a, t, d(t)) = \frac{f_I(a|x, y) f_O(y|x)}{f_I(a|x)},$$

where $x = t - a$ is the year of birth.

Delay time A is the time interval from tumor onset to diagnosis, $A = a - Y$. We will use

A as a surrogate of the natural history of the disease as far as its effect on survival is concerned. A frailty model will be formulated for survival with A as a frailty variable. The effect of early detection due to surveillance is expressed as decreasing A . As a result, the survival function post-diagnosis will be a functional of the distribution of A .

In our first approach we will summarize the effect of A on survival by using the posterior mean delay time as a covariate for survival. Using a specific complete data model for survival, given A , say, a proportional hazards frailty model, may lead to model misspecification. However, if we keep the form of the incomplete data semiparametric survival model flexible, a model building procedure based on the data will absorb such misspecification. This approach is more attractive than traditional frailty modeling, since the choice of a complete data model is difficult to justify anyway, because complete data are not available. Survival methodology developed in our previous project ([Survival Component](#)) specifically addresses flexible model building procedures and guarantees that inference procedures will be available for all models constructed by these procedures.

[Figure_8](#) shows mean posterior delay time (DT) ([Equation Post Y](#)) computed using the marginal incidence model fit to SEER data ([Analysis Of Population Data](#)).

From the [Figure_8](#) it is clear that introduction of PSA is associated with earlier detection, and the older the person the more so. The slight decrease of DT in the no-PSA prediction is a transient process resulting from freezing the pre-PSA trend estimates in year 1988. [Figure_9](#) shows that the DT approach captures the within-stage--and--grade shift. It should be stressed that DT was computed without using survival data and represents an independent significant prognostic factor for post-treatment survival, particularly in localized/regional stage (compare with [Figure_2](#) where a similar effect is expressed by year of diagnosis).

We will proceed as follows.

1. Identify a regression model $f(z|x, y, \xi)$ of stage and grade (z) at diagnosis conditional on independent variables represented by year of birth x , age at tumor onset y , and delay time $\xi = a - y$. This is a regression model for ordered categorical response. We will consider Proportional Odds and Continuation Ratio models. If necessary, custom models will be developed specifically to address this problem if the modelled effect proves to be non-standard. The PI has experience developing models for ordered categorical data and an efficient algorithm for statistical inference with general ordinal models¹.
2. Using the marginal incidence model $f_I(a|x, y)$ [Incidence Model](#), obtain the joint distribution of age and clinical covariates at diagnosis
 $f(a, z|x, y) = f(z|x, y, a)f_I(a|x, y)$. This will contribute to the refined stage- and grade--specific incidence model block that does not use year of diagnosis as a surrogate variable to model stage and grade shift. This block for $f(a, z|x)$ is obtained by integrating out y .
3. Using the joint distribution $f(a, z|x)$, develop a model block for predicting posterior mean age at tumor onset by extending ([Equation Post Y](#)) to include

information on stage and grade at diagnosis. This will improve prediction of DT and the proportion of explained variation in post-treatment survival attributable to the within stage shift.

4. Develop the procedure that will adjust survival time for known screening utilization patterns.

Given a population sample of prostate cancer survival, the procedure will be organized as follows.

1. With the stage- and grade--specific prostate cancer incidence model, obtain the mean posterior DT for each subject in the survival sample.
2. Determine an adequate semiparametric model for survival data with covariates represented by a , $t = x + a$ and mean posterior DT. This would require a trial and error loop through model building using composition techniques (Model Building By Composition), fitting using the Quasi-EM algorithm (Estimation Algorithm) and hypotheses testing using the profile information matrix (Profile Information Matrix).

The significance of our DT approach is that it adjusts survival model for a complex "early detection" confounder. In clinical trials, ignoring significant confounders leads to underestimated treatment effects. With population data, straightforward estimation of the treatment effect is biased since screening utilization is uncontrolled for, and study design is retrospective without randomization. In this project the DT approach serves two main purposes

1. An unbiased assessment of treatment effects with population (and generally retrospective, nonrandomized) data ([Aim Develop Unbiased Assessment Of Treatment Effects](#)).
2. Enables a model that can predict cancer mortality under arbitrary scenarios of utilization of screening and treatment. This paves the way to partitioning mortality into attributable fractions ([Aim Study The Joint Effect Of Progress](#)) short- and long-term predictions of mortality trends ([Aim Make Short And Long Term Predictions](#)), predicting the effects of cancer control strategies that have never been used before, optimization of screening schedules ([Aim Determine Evaluate Optimal Screening Strategies](#)), and addressing other specific aims of this project.

We recognize that mean DT may not provide all of the necessary reduction in the unexplained variation of survival. This will be evaluated by preserving year of diagnosis variable in the model jointly with DT and assessing whether it is still significant. If it turns out that the use of year of diagnosis for explanation of the within stage shift is still necessary, we will extend the DT approach to include the variance of the posterior delay time in addition to the mean and will develop an adjustment of survival using both parameters (mean and variance). Also, as a better but more computer intensive alternative, we will consider using a frailty model approach where posterior distribution for the delay time is used for the frailty variable. The frailty would then represent the uncertainty in the tumor onset given information available at diagnosis in a functional way rather than by one or two surrogate parameters. If



University of Michigan
Delay_Time_Approach
References:

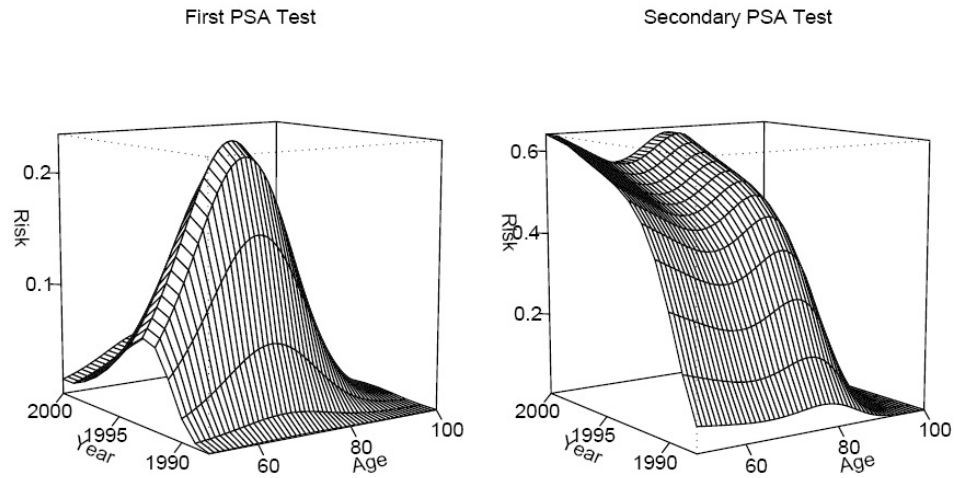
necessary, year of diagnosis trend in addition to the DT-adjustment will be preserved to cover yet unexplained variation of survival.

REFERENCES:

- ¹ Tsodikov, Alexander "A Proportional Hazards Model Taking Account of Long-Term Survivors" in Biometrics 1998; 54: 4: 1508-1516



FIGURE_3B



Figure_3b. Risks of first λ_{1S} and secondary λ_{2S} PSA tests as estimated from the simulation model by age and calendar year. Left: Proportion of never screened men at risk getting their first PSA test. Right: Proportion of men screened at least once getting a secondary PSA test.

REFERENCES:

- ¹ Tsodikov, A., Szabo, A., Wegelin, J. "A population model of prostate cancer incidence" in Statistics in Medicine 2006; 25: 16: 2846-2866

MODELING CANCER DETECTION THROUGH SCREENING-B

Modelling cancer detection through screening



Readers Guide
Model Overview
Assumption Overview
Parameter Overview
Component Overview
Output Overview
Results Overview
Key References

This section (from¹) is devoted to modeling the distribution of potential time to screen-based detection $G_{SDx}(\xi|x, y)$ conditional on the year of birth x and age at tumor onset y . It is a somewhat more detailed version of [Modeling Cancer Detection Through Screening](#).

For an arbitrary individual from the target population, consider the "risk" of getting the first screen in his life. Age at first screen may be regarded as a survival time with the instantaneous risk represented by the hazard function $\lambda_{1S}(a, t)$. Naturally, λ_{1S} depends on age a of the person and the current calendar year t . Generally, it is expected that $\lambda_{1S}(a, t)$ increases in t starting with the year of PSA introduction. As a function of a , it is reasonable to expect that $\lambda_{1S}(a, t)$ is increasing initially while the residual life expectancy is still substantial and then decreasing for very old people. An empirical histogram estimate for $\lambda_{1S}(a, t)$ can be obtained by dividing the number of subjects at the age of a receiving their first screen in year t by the total number of subjects with no evidence of the disease in the (a, t) cell. More precisely, we should count tests in the interval $(t, t + dt)$ and divide by dt , which results in the same estimate for the grouping interval $dt = 1$ year. Note that this estimate is inconsistent unless the data are grouped².

The evolution of an x -birth cohort up to the age of a can be represented as a line connecting points $(\tau, x + \tau)$, where $\tau \in [0, a]$ on the age by year plane called the Lexis diagram³. The probability of no screens by the age of a , G_{1S} , is a survival function obtained by intergrating (accumulating) the hazard λ_{1S} over the line

$$G_{1S}(a|x) = \left\{ - \int_0^a \lambda_{1S}(\tau, x + \tau) d\tau \right\}.$$

Denote by $\lambda_{2S}(a, t)$ the intensity of screening in subjects who already had their first screen. Generally, we expect λ_{2S} to be larger than λ_{1S} . Indeed, the fact that the subject has had his first PSA test may identify him as a member of the group that is screened more frequently for reasons such as easier access to secondary testing having done this once already, favorable attitude towards screening in those who choose to have their first test, doctor's recommendations for serial secondary screens following the first one, etc.

The model for risk of diagnosis by cancer screening is based on the following assumptions.

- The probability that a subject born in year x who has never been screened by the age of a receives his first screen in the age interval $(a, a + da)$ is $\lambda_{1S}(a, x + a)da + o(da)$.
- The probability that a subject born in year x who has been screened at least once by the age of a receives a screen in the age interval $(a, a + da)$ is $\lambda_{2S}(a, x + a)da + o(da)$. This assumption defines secondary screens as following a non-homogeneous Poisson process in age with intensity $\lambda_{2S}(a, x + a)$.

- The probability that a subject born in year x , with the disease onset at the age of y , screened at the age of a is detected with cancer is

$$\begin{cases} 0, & y > a, \\ \alpha(a - y), & \text{otherwise,} \end{cases},$$

where $\alpha(\xi)$ is the sensitivity of screening, and ξ is the age of tumor at the time of testing. It is natural to specify $\alpha(\xi)$ as an increasing function.

It should be noted that the fact that $\lambda_{1S} \neq \lambda_{2S}$ violates the notion that the entire screening schedule for a subject could be a realization of a non-homogeneous Poisson process.

Consider the probability $G_{2SDx}(\tau|x, a, y)$ that a subject born in year x , with onset of the disease at the age of y who has had his first screen by the age of a is not diagnosed by screening in the age interval $[a, a + \tau]$, $a \geq y$. Note that this is a probability of no event in the interval $[0, \tau]$ for a non-homogeneous Poisson process in $\zeta \in [0, \tau]$ with intensity $\lambda_{2S}(a + \zeta, x + a + \zeta)$ thinned with probability $\bar{\alpha}(\zeta + a - y) = 1 - \alpha(\zeta + a - y)$. (We use the notation $\bar{A} = 1 - A$ for any A .) The intensity of a Poisson process with intensity λ thinned with probability $\bar{\alpha}$ is given by the product $\lambda\bar{\alpha}$, so that with $a \geq y$,

$$G_{2Dx}(\tau|x, a, y) = \exp \left\{ - \int_0^\tau \lambda_{2S}(a + \zeta, x + a + \zeta) \alpha(\zeta + a - y) d\zeta \right\}.$$

If the interval in question is before onset, $a + \tau \leq y$, then there is no diagnosis and $G_{2SDx}(\tau|x, a, y) = 1$. If $a < y$ and $a + \tau > y$, the time interval in ζ where diagnosis is possible starts at $y - a$, so that $G_{2SDx}(\tau|x, a, y)$ is given by an expression similar to (\ref{g2dxc}) with the lower limit in the integral set at $y - a$. Summarizing, we have

$$G_{2SDx}(\tau|x, a, y) = \exp \left\{ - \int_{\max(y-a, 0)}^\tau \lambda_{2S}(a + \zeta, x + a + \zeta) \alpha(\zeta + a - y) d\zeta \right\},$$

where $\int_a^b = 0$ for any $b \leq a$.

We are now equipped to derive the probability of no screening diagnosis by the age of $y + \xi$, conditional on year of birth x and age at disease onset y , where ξ is time since onset. We have

$$G_{SDx}(\xi|x, y) = G_{1S}(y+\xi|x) + \bar{G}_{1S}(y|x) G_{2SDx}(\xi|x, y, y) + \int_0^\xi \bar{\alpha}(\nu) f_{1S}(y + \nu|x) G_{2SDx}(\xi - \nu|x, y + \nu, y) d\nu.$$

The first term in the above equation addresses the possibility of no screens by the age of $y + \xi$. The second term addresses the situation when the 1st screen occurs before onset of the disease at the age of y and no diagnosis is achieved through secondary screens that might happen in the age interval $(y, y + \xi)$. The third term accumulates the probability that cancer is missed at the first and secondary screens occurring after disease onset.



REFERENCES:

- ¹ Tsodikov, A., Szabo, A., Wegelin, J. "A population model of prostate cancer incidence" in *Statistics in Medicine* 2006; 25: 16: 2846-2866
- ² Burr, D. "On inconsistency of Breslow's estimator as an estimator of the hazard rate in the cox model" in *Biometrics* 1994; 50: 4: 1142-1145
- ³ Breslow, N. E., Lubin, J. H. , Marek, P., Langholz, B. "Multiplicative models and cohort analysis" in *Journal of the American Statistical Association* 1983; 78: 381: 1-12



University of Michigan
Equation QEM

EQUATION QEM

$$\Delta H_m^{(k+1)} = \frac{D_m}{\sum_{ij \in \mathcal{R}_m} \Theta(F_i^{(k)} | \beta_{ij}, z_{ij}, c_{ij})},$$

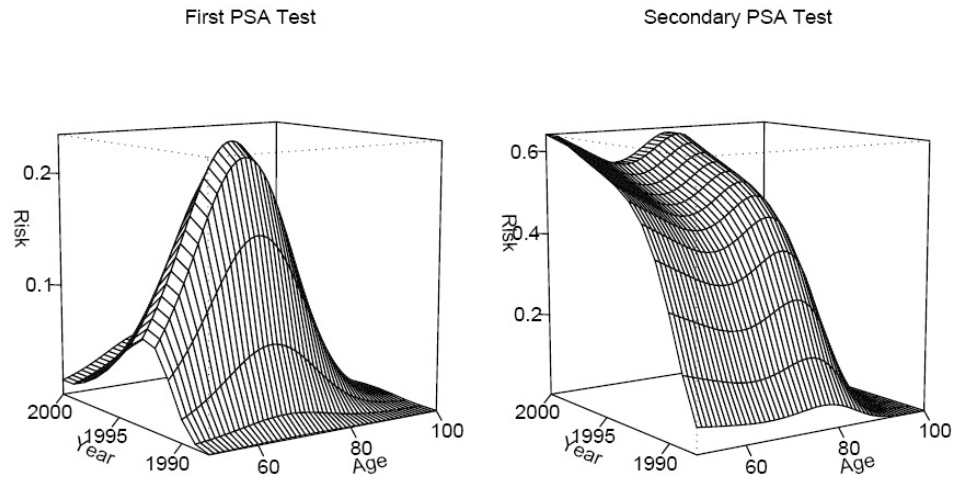


[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)

FIGURE FIRST PSATEST SECONDARY PSATEST



Readers Guide
Model Overview
Assumption Overview
Parameter Overview
Component Overview
Output Overview
Results Overview
Key References

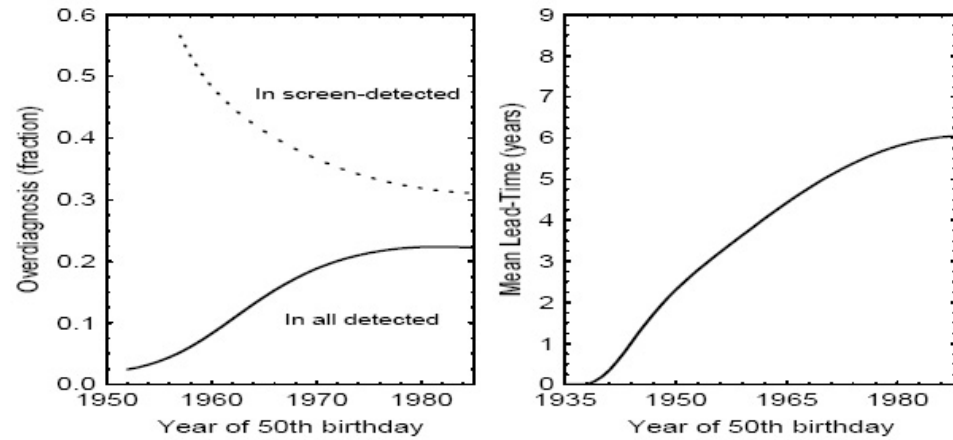


Risks (rate per person) of first λ_{1S} and secondary λ_{2S} PSA tests as estimated from the NIH simulation model by age and calendar year. This simulator is based on data from the National Health Interview Survey (NHIS) and Surveillance, Epidemiology and End Results (SEER) - Medicare linked database (Seer_Medicare). Left: Proportion of never screened men at risk of getting their first PSA test. Right: Proportion of men screened at least once getting a secondary PSA test. Original simulated data were smoothed by a generalized additive model with a logit link.

For more information see [PSAscreening Model](#).



FIGURE_5

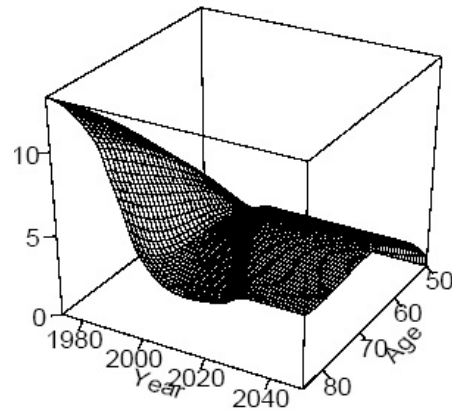


Figure_5. Overdiagnosis (left) and lead-time (right) by birth cohort. Dashed line is the fraction of overdiagnosis in screen-detected patients. Solid line (left) is the fraction of overdiagnosis in all cancer patients.

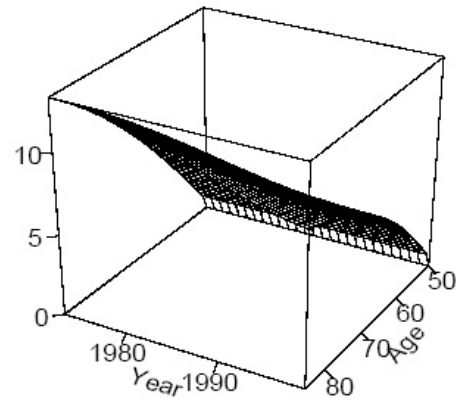


FIGURE_8

(A) PSA, Long-term



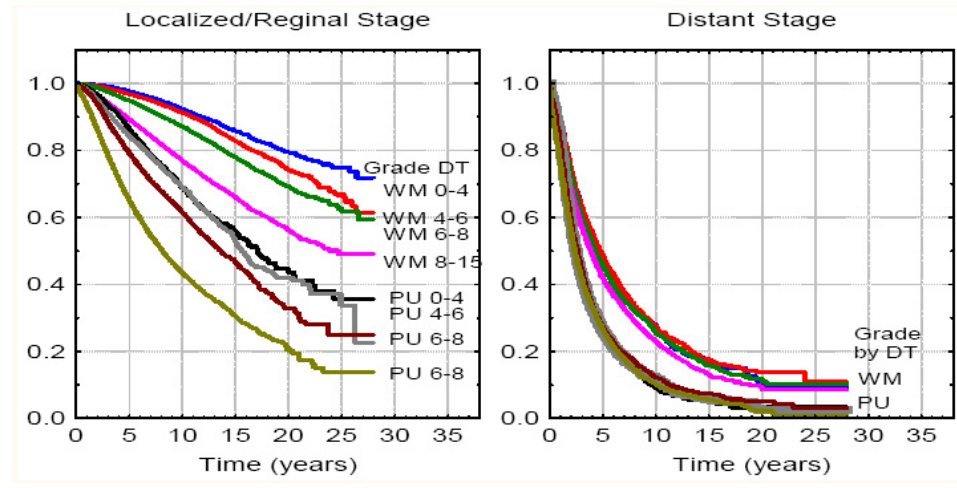
(B) No PSA



Figure_8. Mean posterior delay time by age and year of diagnosis with (A) and without (B) PSA screening. Estimates from SEER data.



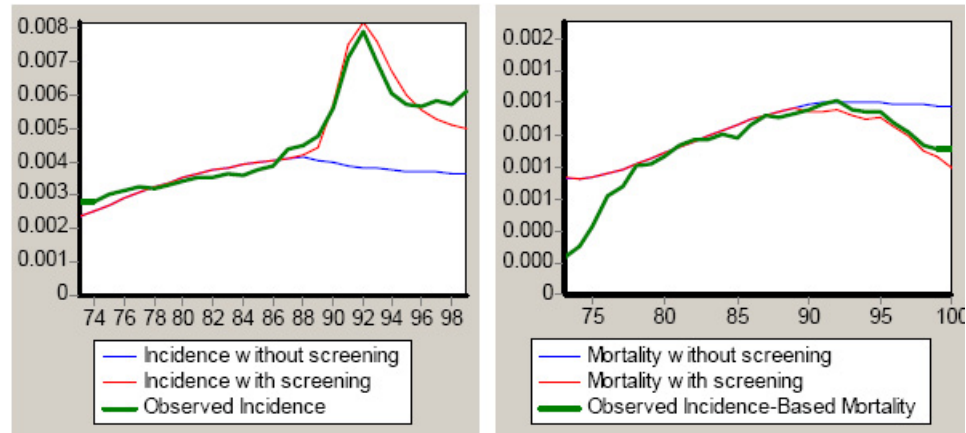
FIGURE_9



Figure_9. Survival by mean posterior delay time (DT), stage (Localized/Regional, Distant) and Grade (Well or Moderately (WM), Poorly or Undifferentiated (PU)). Estimates were obtained from SEER data.



FIGURE_6



Figure_6. Age-adjusted estimates of incidence (left) and mortality (right) in the presence of PSA (red) and prediction of the no-PSA case (blue). Rates are given per person. Thick green curves correspond to empirical estimates. Model-based predictions show overall mortality while empirical estimate is for incidence-based mortality only for cases diagnosed between 1973 and 2000. The discrepancy in the mortality figure for years close to 1973 shows the effect of prostate cancer cases prevalent in 1973.



University of Michigan
Equation LCDX

EQUATION LCDX

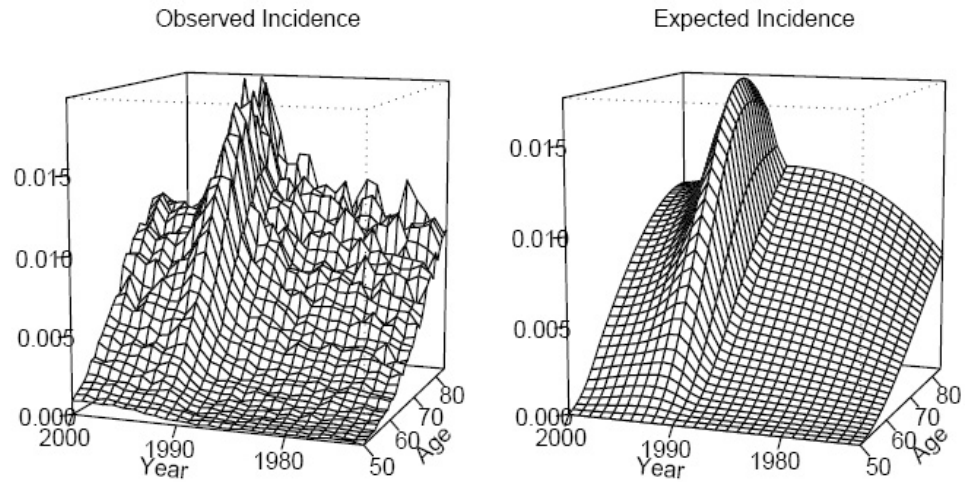
$$\lambda_{CDx}(\xi|x, y) = h_{CDx}(\xi|y)T_{CDx}(x + y + \xi).$$



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)



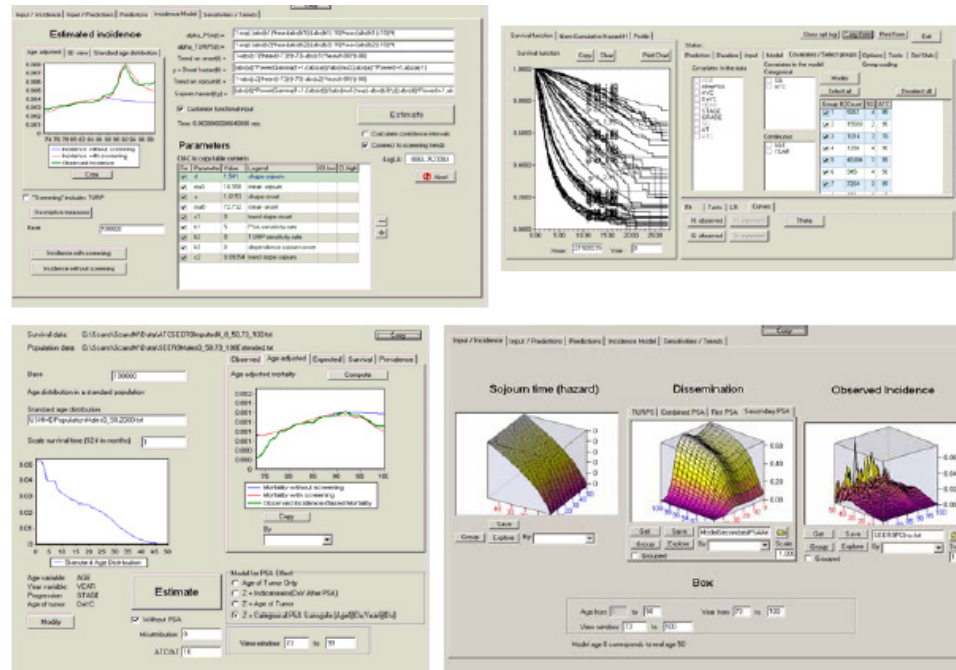
FIGURE_4



Figure_4. Prostate cancer incidence (rate per person). Observed (left): Empirical estimate of prostate cancer incidence computed by dividing incident cancer cases from the SEER database by the population, for each age and calendar year. Expected (right): Model-predicted prostate cancer incidence by age and calendar year.



FIGURE_7



Figure_7. Screen shots of the prototype software package implementing the model.



University of Michigan
Equation Post Y

EQUATION POST Y

$$Y | \{a, t, d(t)\} \propto f_O(y|a, t, d(t)) = \frac{f_I(a|x, y)f_O(y|x)}{f_I(a|x)},$$

where $x = t - a$ is the year of birth.



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)



University of Michigan
Aim Develop Unbiased Assessment Of
Treatment Effects

AIM DEVELOP UNBIASED ASSESSMENT OF TREATMENT EFFECTS

To develop unbiased assessment of treatment effects from population data.



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)



University of Michigan
Aim Study The Joint Effect Of Progress

AIM STUDY THE JOINT EFFECT OF PROGRESS

To study the joint effect of progress in treatment of prostate cancer and PSA utilization on observed national incidence and mortality trends.



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)



University of Michigan
Aim Make Short And Long Term
Predictions

AIM MAKE SHORT AND LONG TERM PREDICTIONS

To make short- and long-term predictions of the trends in national incidence and mortality under various scenarios of projected behavior of key determinants of population processes.



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)



University of Michigan
Aim Determine Evaluate Optimal
Screening Strategies

AIM DETERMINE EVALUATE OPTIMAL SCREENING STRATEGIES

To determine and evaluate optimal screening strategies and predict their effect on future national trends in prostate cancer incidence and mortality.



[Readers Guide](#)
[Model Overview](#)
[Assumption Overview](#)
[Parameter Overview](#)
[Component Overview](#)
[Output Overview](#)
[Results Overview](#)
[Key References](#)



University of Michigan
NHIS

NHIS

REFERENCE

National Center for Health Statistics (2004), "National Health Interview Survey (NHIS)."

URL:

<http://www.cdc.gov/nchs/nhis.htm>



Readers Guide
Model Overview
Assumption Overview
Parameter Overview
Component Overview
Output Overview
Results Overview
Key References

NOTES AND DISCUSSION

CategoryReferences



KEY REFERENCES

- Bolla, M., Gonzalez, D., Warde, P.** (1997) Improved survival in patients with locally advanced prostate cancer treated with radiotherapy and goserelin in *The New England Journal of Medicine* 337 : , p 295–300
- Breslow, N. E., Lubin, J. H. , Marek, P., Langholz, B.** (1983) Multiplicative models and cohort analysis in *Journal of the American Statistical Association* 78 : 381 , p 1-12
- Broet, P., , Tsodikov, A., , De Rycke, Y., , Moreau, T.** (2004) Two-sample statistics for testing the equality of survival functions against improper semi-parametric accelerated failure time alternatives: An application to the analysis of a clinical trial in *Lifetime Data Analysis* 10 : , p 103-120
- Burr, D.** (1994) On inconsistency of Breslow's estimator as an estimator of the hazard rate in the cox model in *Biometrics* 50 : 4 , p 1142-1145
- De Vita, V.T. Jr. , Hellman, S., Rosenberg, S.A.** (1997) Cancer: Principles and Practice of Oncology
- Draisma, G., Boer, R., Otto, S.J., van der Cruijssen, I.W., Damhuis, R.A.M., Schröder, F.H., de Koning, H.J.** (2003) Lead times and overdetection due to prostate-specific antigen screening: Estimates from the European randomized study of screening for prostate cancer in *Journal of the National Cancer Institute* 95 : 12
- Etzioni, R., Cha, R., Feuer, E. J., Davidov, O.,** (1998) Asymptomatic incidence and duration of prostate cancer in *American Journal of Epidemiology* 148 : , p 775-785
- Etzioni, R., Legler, J.M., Feuer, E.J., Merrill, R.M., Chronin, K.A., Hankey, B.F.** (1999) Cancer surveillance series: Interpreting trends in prostate cancer - Part III: Quantifying the link between population prostate-specific antigen testing and recent declines in prostate cancer mortality in *Journal of the National Cancer Institute* 91 : , p 1033–1039
- Feuer, E.J., Merrill, R.M., Hankey, B.F.** (1999) Cancer surveillance series: Interpreting trends in prostate cancer, part ii: Cause of death misclassification and the recent rise and fall in prostate cancer mortality in *Journal of the National Cancer Institute* 91 : , p 1025–1032
- Gamel , J.W., Vogel, R.L.** (2001) Non-parametric comparison of relative versus cause-specific survival in Surveillance, Epidemiology and End Results (SEER) programme breast cancer patients in *Statistical Methods in Medical Research* 10 : , p 339-352
- Hanks, G.E., Pajak, T.F., Porter, A., Grignon, D., Brereton, H., Venkatesan, V., Horwitz, C., Lawton, E.M., Rosenthal, S.A., Sandler, H.M., Shipley, W.U. ,** (2003) Phase iii trial of long-term adjuvant androgen deprivation after neoadjuvant hormonal cytorreduction and radiotherapy in locally advanced carcinoma of the prostate in *Journal of Clinical Oncology*, 21 : , p 3972–3978
- Himmelblau, D.M.** (1972) Applied Nonlinear Programming
- Hoffman, R.M., Stone, S.N., Hunt, W.C., Key, C.R., Gilliland, F.D.** (2003) Effects of misattribution in assigning cause of death on prostate cancer mortality rates in *Annals of Epidemiology* 13 : , p 450-454
- Holmberg, L.H., Bill-Axelson, A., Helgesen, F., Salo, J.O., Folmerz, P., Haggman, M., Andersson, S.-O., Spangberg, A., Busch, C., Nordling, S., Palmgren, J., Adami, H.-O., Johansson, J.-E., Norlen, B.J.** (2002) A randomized trial comparing radical prostatectomy with watchful waiting in early prostate cancer in *The New England Journal of Medicine* 347 : , p 741-789
- Hougaard, P.** (1996) Frailty models for survival data in *Lifetime Data Analysis* 1 : , p 255–274

- Hsing, A.W., Devesa, S.S.** (2001) Trends and patterns of prostate cancer: What do they suggest? in *Epidemiologic Reviews* 23 : , p 3-13
- Hu, P., Zelen, M.** (1997) Planning clinical trials to evaluate early detection programmes in *Biometrika* 84 : , p 817-829
- Lawless, J.F.** (1987) Regression methods for Poisson process data in *Journal of the American Statistical Association* 82 : , p 808-815
- Lawton, C.A., Winter, K., Murray, K., Machtay, M., Mesic, J.B., Hanks, G.E., Coughlin, C.T., Pilepich, M.V.** (1998) Updated results of the phase iii radiation therapy oncology group (rtog) trial 85- 31 evaluating the potential benefit of androgen suppression following standard radiation therapy for unfavorable prog in *International Journal of Radiation Oncology, Biology and Physics* 93 : , p 937-46
- Lee, S.J., Zelen, M.** (2002) Statistical models for screening: planning public health programs in *Cancer Treat Res* 113 : , p 19-36
- Lee, S.J., Zelen, M.** (1998) Scheduling periodic examinations for the early detection of disease: applications to breast cancer in *Journal of the American Statistical Association* 93 : , p 1271-1281
- Merrill, R.M., Feuer, E. J., Warren, J. L. , Schussler, N. , Stephensons, R. A.** (1999) Role of transurethral resection of the prostate in population-based prostate cancer incidence rates in *American Journal of Epidemiology* 150 : 8 , p 848-860
- Merrill, R.M., Wiggins, C. L., ,** (2002) Incidental detection of population-based prostate cancer incidence rates through transurethral resection of the prostate in *Urologic Oncology* 7 : , p 213-219
- Messing, E.M., Manola, J., Sarosdy, M., Wilding, G., Crawford, E.D., Trump, D.,** (1999) Immediate hormonal therapy compared with observation after radical prostatectomy and pelvic lymphadenectomy in men with node-positive prostate cancer in *The New England Journal of Medicine* 341 : , p 1781-1788
- National Cancer Institute** (2002) Surveillance Epidemiology and End Results (SEER) - Medicare linked database
- National Center for Health Statistics** (2004) National Health Interview Survey (NHIS).
- Optenberg, S. A., Thompson, I. M., Friedrichs, P., Wojcik, B., Stein, C. R., Kramer, B.** (1995) Race, treatment, and long-term survival from prostate cancer in an equal-access medical care delivery system in *Journal of the American Medical Association* 274 : , p 1599-1605
- Potosky, A.L., Feuer, E.J. , Levin, D.L.** (2001) Impact of screening on incidence and mortality of prostate cancer in the United States in *Epidemiological Review* 23 : 1 , p 181-186
- Roach III, M., Lu, J., Pilepich, M. V., Asbell, S. O., Mohiuddin, M., Grignon, D.** (2003) Race and survival of men treated for prostate cancer on radiation therapy oncology group phase iii randomized trials in *Journal of Urology* 169 : , p 245-250
- Roach, M., III, Krall, J., Keller, J. W., Perez, C. A., Sause, W. T., Doggett, R. L., Rotman, M., Russ, H., Pilepich, M. V., Asbell, S. O.** (1992) The prognostic significance of race and survival from prostate cancer based on patients irradiated on Radiation Therapy Oncology Group protocols (1976-1985) in *International Journal of Radiation Oncology Biology and Physics* 24 : , p 441-449
- Robbins, A. S., Whittemore, A. S., Van Den Eeden, S. K.** (1998) Race, prostate cancer survival, and membership in a large health maintenance organization in *Journal of the National Cancer Institute* 90 : , p 986-990
- Sasieni, P.D.** (1996) Proportional excess hazard in *Biometrika* 83 : , p 127-141

- Scardino P.T.** (1989) Early detection of prostate cancer in *Urol Clin North Am* 16 : , p 635-655
- Shen, Y., Zelen. M.** (1999) Parametric estimation procedures for screening programmes: Stable and nonstable disease models for multimodality case finding in *Biometrika* 86 : , p 503-515
- Stewart, S.L., King, J.B., Thompson, T.D., Friedman, C., Wingo, P.A.** (2004) Cancer mortality surveillance--United States, 1990-2000 in *MMWR Surveillance Summary* 53 : 3 , p 1-108
- Thompson, I., Tangen, C., Tolcher, A., Crawford, E., Eisenberger, M., Moinpour, C.** (2001) Association of african-american ethnic background with survival in men with metastatic prostate cancer in *Journal of the National Cancer Institute* 93 : , p 219-225
- Thompson, I.M., Goodman, P.J., Tangen, C.M., Lucia, M.S., Miller, G.J., Ford, L.G., Lieber, M.M., Cespedes, R.D., Atkins, J.M., Lippman, S.M., Carlin, S.M., Ryan, A., Szczepanek, C.M., Crowley, J.J., Coltman, C.A.** (2003) The influence of finasteride on the development of prostate cancer in *The New England Journal of Medicine* 349 : , p 215-224
- Tsodikov, A., Garibotti, G.** (2007) Profile information matrix for nonlinear transformation models in *Lifetime Data Analysis* 13 : 1 , p 139-159
- Tsodikov, A., Ibrahim, J.G., Yakovlev, A.Y.** (2003) Estimating cure rates from survival data: An alternative to two-component mixture models in *Journal of the American Statistical Association* 98 : , p 1063-1078
- Tsodikov, A.** (2002) Semiparametric models of long- and short-term survival: An application to the analysis of breast cancer survival in utah by age and stage in *Statistics in Medicine* 21 : , p 895-920
- Tsodikov, A.** (2006) Using composition to build semiparametric survival models in *Statistical Modelling* Submitted :
- Tsodikov, A., Szabo, A., Wegelin, J.** (2006) A population model of prostate cancer incidence in *Statistics in Medicine* 25 : 16 , p 2846-2866
- Tsodikov, Alex** (2003) Semiparametric models: a generalized self-consistency approach in *Journal of the Royal Statistical Society, Series B* 65 : , p 759-774
- Tsodikov, Alexander** (1998) A Proportional Hazards Model Taking Account of Long-Term Survivors in *Biometrics* 54 : 4 , p 1508-1516
- Weller, E.A., Feuer, E.J., Frey, C.M., Wesley, M.N.** (48) Parametric relative survival regression using generalized linear models with application to Hodgkin's lymphoma in *Applied Statistics* 1999 : , p 79-89
- Wilmoth, John R. (Director), Shkolnikov, Vladimir Shkolnikov (Co-Director),** (2003) Human Mortality Database (HMD).
- Wood, S.** (2003) Thin plate regression splines in *Journal of the Royal Statistical Society Series B* 65 : , p 95-114
- Woolf, S.H.** (1995) Screening for prostate cancer with prostate-specific antigen. in *he New England Journal of Medicine* T333 : , p 1401-1405
- Zelen, M., Feinleib, M.** (1969) On the theory of screening for chronic diseases in *Biometrika* 56 : , p 601-614

Core Docs