Screening and Health Transition

Yun Zhang\textsuperscript{a}, Donna B. Gilleskie\textsuperscript{b} and John T. Wei\textsuperscript{c}

\textsuperscript{a}Institute for Healthcare Policy and Innovation, University of Michigan at Ann Arbor
\textsuperscript{b}Department of Economics, University of North Carolina at Chapel Hill
\textsuperscript{c}Department of Urology, University of Michigan at Ann Arbor

March 21, 2013

SUMMARY

We estimate individual health evolution using a principal-agent model that incorporates health screening results but does not depend on health screening policy at data collection. Our model characterizes age-specific health state transitions as functions of diagnosis history (e.g., earliness of detection and level of diagnosed disease) and allows for patient non-compliance with screening recommendation. Using patient survey data and medical claims data, we circumvent two challenges in screening policy assessment: restricted applicability associated with clinical trial data and the extrapolation concern created by prohibitively costly long-term patient follow-up. We apply our approach to evaluation of prostate cancer screening on baby boomers life expectancies.

Keywords: screening, latent health transition, principal-agent problem, Markov process.

JEL Code: C53, C54, C18, I18, I12, D04
1 Introduction

Due to major advances in both medical imaging technology and human genome research over the past two decades, new screening tests for detecting common disorders such as heart disease, diabetes, and cancer are expanding rapidly (Maitino et al., 2003; Khoury et al., 2003). Screening practices trigger earlier diagnosis of disease and therapeutic intervention which, in turn, influence health transition. Unlike typical medical care interventions, a screening program targets a general asymptomatic population rather than the symptomatic sub-population (Wald, 2008). Therefore, screening policies have major public health implications in terms of both effectiveness of care and its costs.

Screening, which affects diagnostic test utilization, indirectly influences health transition by perturbing theearliness of treatment over a multiple-period time horizon. Due to the complex clinical pathway from diagnostic test use to final outcomes, methodological challenges associated with assessing diagnostic test policy have been well recognized for decades in the health technology assessment community (Tatsioni et al., 2005; Bossuyt et al., 2006; den Bruel et al., 2007). Since the widely cited report by Wilson and Jungner (1968), numerous studies have attempted to evaluate the impact of screening on various disease outcomes with findings that resulted in conflicting policy recommendation (Gazelle et al., 2005; Pearson et al., 2008). Consider, for example, prostate cancer, which is the most prevalent solid tumor cancer among American men. The U.S. Preventive Services Task Force (USPSTF) recommends against Prostate-Specific Antigen (PSA) screening for prostate cancer in men (USPSTF, 2011a); in contrast, the American Urologic Association recommends consideration of PSA screening among men age 50 and above (AUA, 2011). Similarly, the USPSTF and American Cancer Society have announced conflicting recommendations (USPSTF, 2011b; ACS, 2011) for the screening of breast cancer, the most common cancer among American women. These screening policy controversies have sparked heated on-going national debates (Barry, 2009; Hobson, 2010; Brownlee and Lenzer, 2011; Parker-Pope, 2012).  

1It is worth noting that the duration of the multiple-period itself is an endogenous quantity because screening affects survival.  

2The majority of these debates center around whether screening improves health outcomes (Gazelle et al., 2005; Woolf and Harris, 2012); issues related to costs are not the focus. The appropriateness of a particular screening policy becomes even more controversial when the decisionmaker introduces other considerations such as social payoffs and opportunity costs.
To date, few studies have provided a generic theoretical framework that characterizes optimal screening behavior on the basis of exogenous determinants such as disease prevalence or the natural history of disease, test accuracy, and treatment efficacy (Wilson and Jungner, 1968). Likewise, few economists have examined the issue of screening despite its natural inclusion in economic discussions ranging from diminishing marginal returns, to principal-agent relationships, to policy invariant model specifications (i.e., the Lucas critique). The seminal work by Phelps and Mushlin (1988) (henceforth referenced as the PM model) is an exception. The unique advantage of the PM model stems from the ability to solve for the optimal diagnostic test policy inside a social payoff maximization problem. Unfortunately, the adoption of the PM model in screening research has been limited (Cantor et al., 1999; Hilden, 2004; Laking et al., 2006) by its two restrictive assumptions. First, it assumes that a patient takes a diagnostic test only once. A patient, in general, may take a diagnostic test repeatedly over his lifecycle. Second, a diagnostic test usually includes both a presumptive test and a confirmatory test in sequence. However, informed screening policy assessment requires measurement of the effects of both the former test and the latter test on health outcomes.

Given the recent call for screening policy evaluation that uses comparative effectiveness analysis (Wilensky, 2006), we develop a screening policy decision framework that (1) characterizes the impact of the determinants of screening (as documented in Wilson and Jungner, 1968) on health evolution over a life cycle and is not bound by restrictive assumptions, and (2) allows for derivation of an optimal screening policy function and social benefit frontier of screening based on the comparative effectiveness standard.

In addition, our model addresses two common data issues that challenge empirical screening research. First, existing data sources provide little variation in the components of a screening policy (e.g., age of first use, testing criteria, etc.). Impediments to variation

---

3The following works discuss some aspects of screening using an economics approach, but do not provide a comprehensive model of health transitions: Meltzer (2001), Gyrd-Hansen et al. (1998), Whynes et al. (1998), Gerard et al. (2003), Laking et al. (2006), Marshall et al. (2009), and Stollenwerk et al. (2012). For example, the long-standing controversy over prostate cancer screening focuses on PSA, a presumptive test.

5By specifying a costing function based on the same domain of health transition functions in the model, optimal screening policy that maximizes the social benefit specification in the PM model can be derived. Details of this derivation are available from the authors upon request.
are numerous. For any given disease, at most only a few screening interventions have been evaluated at a population level due to high experimental costs. While observational data (e.g., medical claims) often records diagnostic test utilization, the variation in available tests is limited. Even if the variation is sufficient to estimate the mortality rate as a function of the screening policy, the reduced-form model is of limited value if one wishes to evaluate new interventions of interest that may lie beyond the range of the policies available in the data used for estimation (i.e., due to concern about inappropriate model extrapolation).

Second, the data requirements for evaluation of screening policy are demanding, especially for screening intervention studies. To estimate screening’s effects on life expectancy using a typical reduced-form analysis, one needs to follow study subjects over their entire post-intervention life cycles. At a minimum, researchers must follow patients for a specified time period in order to calculate survival rates for that duration. This requirement hinders the timeliness of policy decision because the length of time required for registering death events, or assessing duration-specific survival rates, is substantial.

To address these two data issues, health evolution in our model incorporates the information gained from health screening but is not dependent on the specifics of any health screening recommendation itself. Our model characterizes age-specific health state transitions as functions of the history of diagnosis. Specifically, the transitions depend on the first period of detection of disease and the level of diagnosed disease. As long as the number of observations that populate each age-health state combination is sufficiently large, it is not necessary that the data sample follow subjects over their entire post-intervention life cycle. Estimation of our model does require the ability to trace health histories. This requirement is, in general, much less demanding than following subjects over their life cycles. We find

---

6For example, prostate cancer screening tests using PSA levels are based on a small number of PSA cut-off values (e.g., 2.5 ng/ml and 4 ng/ml) to indicate an abnormality.

7The typical reduced-form survival model for assessing screening’s effect on a cohort of subjects specifies an indicator of death (or censoring) as its dependent variable at each time period and the screening policy as the explanatory variable. If study participants are followed from age 65 for, say, ten years, then it is difficult to estimate the mortality hazard after age 75 without relying on strong assumptions. Policy makers are often interested in assessing the lifetime effect of screening on relatively young patients (e.g., age 50). Hence, empirical approaches that could lessen this reliance on lengthy subject follow-up are needed.

8For example, the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO) is a large-scale clinical trial designed to determine whether certain cancer screening tests reduce deaths from cancer in these sites. The trial began in 1992 and screening of participants ended in 2006. It is believed that at least ten years of follow-up after 2006 is needed to fully register the long-term outcomes. (Source: http://prevention.cancer.gov/plco)
that at most eight years of health state histories for prostate cancer patients at each age were needed to model health transition. That is, the estimation results provide evidence that the marginal effect of an early diagnosis is diminishing over time: conditional on age, current health state, and the diseased state at the time of detection, the length of time post-diagnosis does not provide any additional impact on health transitions after five to eight years.

Our approach differs from existing screening models in several important ways. (See the review by Stevenson (1995) and Wever et al. (2011).) Our model incorporates patient preferences in order to explore non-compliance with screening recommendations. Our model recovers the survival history, natural disease history, and screening history based on the same health transition functions that capture the benefit/harm of early diagnosis. Furthermore, our model exploits the observed (in the data) timing of diagnosis and diagnosis disease severity that differ across different disease screening policies in order to measure the early diagnosis effect. Such a design ensures that health transition functions are (1) primitive enough to maintain their independence from screening policy (valuable for solving extrapolation issues), and (2) not overly primitive at the cost of the model’s applicability across diseases. Lastly, our model addresses the selection bias (e.g., healthy-screenee effect) that arises from estimation using observational data (Zeliadt et al., 2007).

The remainder of the paper is organized as follows. Section 2 introduces the theoretical model where optimal screening policy is derived from a multiple time period principal-agent problem with full commitment. It is the earliness of disease diagnosis rather than a particular screening intervention that directly affects the health transition process. Based on this insight, in the model, we specify age-specific health transition functions in terms of timing of diagnosis and diagnosed disease state. We show that our model maintains the useful property that screening policy and health evolution are independent and we characterize

9Screening models that estimate parameters of disease-specific cellular biology (e.g., growth of cancer cells) are called ‘deep’ models (Stevenson, 1995). A model with its parameters ‘deep’ enough to be constant across screening policies is one that avoids concern associated with policy extrapolation. However, if the model is so ‘deep’ that its measurements of earliness of diagnosis are disease-specific then the generalization of the model is limited. An ideal model meets two conditions: (1) its earliness of diagnosis measurements are easily available, and (2) its health transition functions are screening policy invariant.

10A game with full commitment means that the principal (i.e., the screening policymaker) will fully commit to screening recommendations for the study time horizon chosen at the beginning of the study time horizon.
optimal screening policy and the frontier of screening’s social benefit based on the comparative effectiveness criterion. We show that under two weak assumptions about a confirmatory diagnostic test – (1) the test is perfectly specific and (2) the test correctly ascertains the disease state for a true positive patient – and using information on observed medical care treatment, the researcher can estimate the model with ease. In Section 3, we apply our model to prostate cancer disease in order to assess the effect of PSA screening on the life expectancy of elderly American men using nationally-representative claims data and retrospective data. Section 4 presents the simulated health transitions associated with various screening recommendation scenarios using the estimated model. Section 5 concludes with a discussion of model advantages and limitations.

2 Theoretical Model

In this section, we present a model for comparative effectiveness assessment of screening in which optimal screening decision arises from a multiple-period principal-agent problem with full commitment. In the model, individual health evolution incorporates the information gained from health screening but is independent from the specifics of any health screening recommendation itself. The health transitions are captured by a high-order Markov process that, when fit to data from a random sample of representative individuals, recovers parameters that are invariant to any screening policy that may exist at the time of data collection. The model describes latent (unobserved) health over time; individual data detail screening outcomes and observed (or diagnosed) health.

This section begins with a description of health states and the diagnostic tests used in screening. Next, it characterizes individual health evolution as a high-order Markov process and proves policy-invariance of the individual health evolution process. Afterwards, it lays out the principal-agent problem in which the policy maker maximizes social benefit over the population based on the comparative effectiveness criteria. The section ends with a detailed discussion of strategies for estimating the model using observational data.
2.1 Latent health states and their initial distribution

The theoretical model considers a representative individual/patient observed over a finite time horizon for \( J \) periods, \( t = t_0, t_1, \ldots, t_J \), starting from age \( t_0 \). In the model, the individual’s latent health status in period \( t \) is described by two independent health states, \( h_t^* \) and \( \bar{h}_t^* \). The first health state, \( h_t^* \), for any period \( t \) takes on values defined by the time-invariant set \( H^* \equiv \{ h_1^*, h_2^*, \ldots, h_n^* \} \). These \( n \) categorical outcomes encompass the possible states of latent (unobserved) health that may be detected by a diagnostic test of policy interest (e.g., the PSA screening test for men). The \( n \) elements of the set \( H^* \) are ordinal; \( h_1^* \) represents a health state “free of the disease of interest” and \( h_n^* \) represents the health state of “death caused by the disease of interest”. Note that this metric is not the outcome of the test itself, but the underlying health state that a test is meant to measure. We are fully aware that a diagnostic test may report the underlying health accurately, or may produce false positive or false negative errors.\(^{11}\) We define the outcome of a screening test later.

The second health state used to describe latent health, \( \bar{h}_t^* \), takes on values in an analogous set \( \bar{H}^* \equiv \{ \bar{h}_1^*, \bar{h}_2^*, \ldots, \bar{h}_\pi^* \} \) that is meant to reflect one’s health as related to all diseases except the disease of interest. Similarly, the \( \pi \) elements of the set \( \bar{H}^* \) are ordinal; \( \bar{h}_1^* \) represents the health state of “being free of any disease other than the disease of interest” and \( \bar{h}_\pi^* \) represents a health state of “death by a disease other than the disease of interest”. We define the two states to reflect the fact that health is a multi-dimensional concept, but we only focus on that part of the health process related to the disease of policy interest (i.e., screening recommendations). To facilitate discussion, we refer to the first set of possible health states as (latent) policy-relevant health states and the second set of possible health states as (latent) policy-irrelevant health states throughout the rest of this paper.\(^{12}\)

\(^{11}\)A false positive is a result (from a diagnostic test) that indicates that a condition of interest is present when it is not. In statistical terms, a type I error is analogous to a false positive result. A false negative is the result that indicates that a condition of interest is not present when it truly is. A type II statistical error is analogous to a false negative result.

\(^{12}\)Policy-irrelevant states are irrelevant to screening policy only in the sense that they cannot be detected by the specific diagnostic test of policy interest. We explain later how we allow the possible components of a screening policy to influence the distribution of policy-irrelevant health indirectly.
2.2 Presumptive diagnostic test and confirmatory diagnostic test

A researcher cannot observe an individual’s latent policy-relevant health state. What he can observe is the diagnostic test result of a patient who is screened. Before defining notation regarding testing outcomes and diagnosed disease state, we discuss the typical process of disease detection undertaken by physicians. A typical disease screening consists of a presumptive diagnostic test (with a positive or negative outcome) and, if this test is positive for disease, a confirmatory diagnostic test (that produces a measured level of disease). The presumptive test (e.g., a blood sample test) is typically used before the more invasive and definitive confirmatory test (e.g., a biopsy, an imaging test, or their combination). Whether or not a patient would be offered the confirmatory test depends on his presumptive test result. If the presumptive test result is negative then the person would not proceed to the confirmatory test; in this case, the patient’s diagnosed disease state is “disease free”. Otherwise, the patient would take the confirmatory test. For a patient who takes the confirmatory test, the person’s final disease indicator and diagnosed disease state would then be based on the confirmatory test results. Typically, screening is synonymous with a presumptive diagnostic test.

The results of a diagnostic test taken in period $t$ consist of two elements, $s^d_t$ and $h^d_t$. The first element is a binary indicator of the presumptive test outcome. If a patient’s disease indicator at time period $t$ is positive (i.e., $s^d_t = 1$), then the diagnosis process would further ascertain his disease state ($h^d_t$)\footnote{Also, once a test at time $t$ indicates disease, the disease state in all subsequent periods is assumed positive. This assumption is formally defined below after introducing some additional notation.}. Unlike the latent policy-relevant health state ($h^*_t$), both the binary disease indicator ($s^d_t$) and the (diagnosed) disease state ($h^d_t$) may not reflect the true health state. Let $H^d$ denote the set of possible (diagnosed) disease states whose categorical values are the same as the $n$-valued set $H^*$. We make the notational distinction because $h^d_t$ is meant to indicate an outcome that is a best guess while the $h^*_t$ counterpart is meant to indicate a latent true state.

To summarize, the health of a patient who has elected to undergo screening can be described by the latent health state and the observed diagnosis state. More specifically,
upon initial observation (in the data), $t_0$, the patient is endowed with an initial policy-relevant health state ($h_{t_0}^*$), an initial policy-irrelevant health state ($\overline{h}_{t_0}$), an initial disease indicator ($s_{t_0}^d$), and an initial diagnosed disease state ($h_{t_0}^d$). The researcher only observes a $K$-dimensional vector of the patient’s exogenous characteristics ($z \in \mathbb{R}^K$), the initial disease indicator, and the initial diagnosed disease state. The researcher knows the (conditional on $z$) density functions of the patient’s initial policy-relevant health states, $F_z(h_{t_0}^*)$, and initial policy-irrelevant health states, $F_z(\overline{h}_{t_0})$, but does not know the true health states themselves. The evolution of a patient’s true health is described by two processes, one pertaining to the policy-relevant health state and the other to the policy-irrelevant health state. We assume that these two processes progress independently prior to the onset of death. When death occurs in either one of the two processes, health transition terminates in both.

### 2.3 Individual health evolution process

In what follows, we begin by characterizing the distributions of outcomes of diagnostic testing. Next, we describe the role of a screening policy in determining the timing of diagnostic testing. Afterwards, we first characterize the evolution of policy-relevant health and then summarize the entire evolution process including policy-irrelevant health.

#### 2.3.1 Diagnostic testing outcomes

Both a single diagnostic test or a diagnostic test sequence consisting of a presumptive test and a confirmatory test is a stochastic mapping from a latent policy-relevant health state to a diagnostic output. The output from a presumptive test is a binary disease indicator of whether or not one has the disease in question. The diagnostic output from a confirmatory

---

14If a patient has not been screened, then the disease indicator is degenerately zero (indicating that no test has been performed) and the disease state is assumed “free of disease”.

15An example of screening policy is the recently released new screening recommendation for the prevention and early detection of cervical cancer (Saslow et al., 2012). The latest recommendations include: all women should begin screening at age 21; women between the ages of 21 and 29 should have a Pap test every 3 years and they should not be tested for HPV unless there is an abnormal Pap test result; women between the ages of 30 and 65 should have both a Pap test and HPV test every 5 years; women over age 65 who have had regular screening with normal results should not be screened for cervical cancer; and additional recommendations for women who have had a positive Pap or HPV test.
test (or a test sequence that ends in a confirmatory test) is more informative because it measures the level, or seriousness, of the disease.

The accuracy of a confirmatory test is summarized by a $n \times n$ confusion matrix $\psi^c$. In the matrix, the element in the $i^{th}$ row and the $j^{th}$ column is the probability that a person with policy-relevant health state, $h^*_i$, is diagnosed with disease state, $h^d_j$. For example, $\psi^c_{2,3}$ is the probability that a person residing in the second policy-relevant health state ($h^*_2$) is diagnosed with the third policy-relevant health state ($h^d_3$). More specifically,

$$
\psi^c = \begin{pmatrix}
\psi^c_{1,1} & \psi^c_{1,2} & \cdots & \psi^c_{1,n} \\
\vdots & \vdots & \ddots & \vdots \\
\psi^c_{n,1} & \psi^c_{n,2} & \cdots & \psi^c_{n,n}
\end{pmatrix}
$$

We have suppressed the time (or age) subscripts above, but the accuracy of a test may be time (or age) dependent. Each row of the confusion matrix sums to one. If a confirmatory test is perfectly accurate, the diagonal elements of its confusion matrix are equal to one.

Given this information, we can now formally define a confirmatory test. Consider a latent policy-relevant health state corresponding to the $i^{th}$ row of the confusion matrix. The categorical output of a confirmatory test would follow a distribution defined by the $n$ probabilities in the $i^{th}$ row of the confusion matrix ($\psi^c_{i,\cdot}$). Let the function $\chi^c$ be a mapping from the true, unobserved health state to the diagnosed health state; that is, $\chi^c : H^* \rightarrow H^d$ denotes a confirmatory test. Hence,

$$
h^d_i = \chi^c(h^*_i) \sim \text{Categorical}(\psi^c) \quad (1)
$$

where $h^d_i \in H^d$. A presumptive test only generates a binary disease indicator. So, its confusion matrix, $\psi^p$, has two columns such that

$$
\psi^p = \begin{pmatrix}
\psi^p_{1,1} & 1 - \psi^p_{1,1} \\
\vdots & \vdots \\
\psi^p_{n,1} & 1 - \psi^p_{n,1}
\end{pmatrix}
$$

16 The rows of a confusion matrix represent the possible true health states, while the columns represent the possible diagnosed disease states. This matrix describes whether the diagnosis procedure is confusing two states.
where $\psi_{p,1}^*$ is the specificity of the presumptive test and $1 - \psi_{p,1}^*$ is the sensitivity of the presumptive test for the $n - 1$ latent policy-relevant health states with disease (i.e., $h_i^*$, $i = 2, \ldots, n$)\(^{17}\)

Similarly, a presumptive test is a stochastic mapping, $\chi^p : H^* \rightarrow \{0, 1\}$, whose output has a binary distribution defined by the confusion matrix $\psi_p$. That is,

$$s^d_t = \chi^p(h_t^*) \sim \text{Binary}(\psi^p) \quad (2)$$

where $s^d_t \in \{0, 1\}$ and “0” corresponds to a diagnosis of no disease and “1” presumes the disease.

For completeness, we bring both of the functions above together to produce a diagnostic test sequence function that defines both the presumptive test and confirmatory test outcomes. Let $\psi \equiv \{\psi^p, \psi^c\}$ and let vector function $\Delta(h_t^*; \psi)$ denote such a diagnosis test sequence for each possible value of the true, unobserved health state ($h_t^*$). More specifically,

$$\Delta(h_t^*; \psi) = [\Delta^p(h_t^*; \psi^p), \Delta^c(h_t^*; \psi^c)]^T = \begin{cases} 
[0 \ h^d_t] & \text{if } \chi^p(h_t^*) = 0 \\
[1 \ \mathbf{X}(h_t^* \in H^d \setminus h^d_t)] \ \mathbf{X}(h_t^*) & \text{if } \chi^p(h_t^*) = 1
\end{cases} \quad (3)$$

where $\Delta^p(\cdot)$ generates a binary disease indicator and $\Delta^c(\cdot)$ generates the categorical (diagnosed) health state for every possible latent health state; $\chi^p(h_t^*)$ and $\chi^c(h_t^*)$ are the presumptive test function and the confirmatory test function, respectively; and $1 \ [\cdot] \ |$ is an indicator function\(^{18}\).

### 2.3.2 Screening policy and individual test participation

Components of a screening policy are those variable features that are “pinned down” by a universal recommendation. A recommendation may specify age, frequency, health history, duration, and types of tests. For example, a screening policy such as that for cervical cancer (see footnote 14) provides age-, test-, and health history-specific guidelines for participation in a diagnostic test sequence. In our model, we assume that an individual’s participation depends on the screening policy, his current age ($t$), and his previous period binary disease indicator ($s^d_{t-1}$). Let $p_t \in \{0, 1\}$ denote the participation decision; $p_t = 1$ means having

\(^{17}\)Specificity is the proportion of negatives that are correctly identified by the diagnostic test, or the true negative rate. Sensitivity is the proportion of positives which are correctly identified, or the true positive rate.

\(^{18}\)Henceforth, we use $[\cdot]_k$, $[\cdot]_{[k]}$ and $[\cdot]_{[k]}$ to denote the $k^{th}$ element of a vector, $k^{th}$ row of a matrix, and $k^{th}$ column of a matrix, respectively.
the diagnostic test(s) (i.e., being screened) and \( p_t = 0 \) means not having it. Individual characteristics \( z \) may influence the participation decision (see below).

Consider a screening policy where age is the only component of recommendation. Over the \( J + 1 \) periods, there are \( 2^{J+1} \) possible screening policies, reflecting the recommendation (or not) of age \( t \) as a screening age. That is, one policy may recommend screening every year after age \( t = t_0 \) and another policy may recommend screening every three years after age \( t \). Obviously, there is no need to be screened if the true health state is death (\( h_t^* = h^*_n \)). We denote any particular screening policy by \( \theta \in \Theta \equiv \mathcal{P}(\{t_0, t_1, \ldots, t_J\}) \cup \text{NA} \). \( \mathcal{P}(\{t_0, t_1, \ldots, t_J\}) \) is the power set of all possible screening periods and \( \text{NA} \) represents no screening recommendation in all periods.\(^{19}\)

Let \( \varrho_z(s_{l-1}, t, h_t^*, \overline{h}_t^*; \theta) \) denote the individual’s test participation function in the presence of a screening recommendation \( \theta \). Because an individual may voluntarily take a test in the absence of any screening policy, we let \( \varrho^v_z(s_{l-1}, t, h_t^*, \overline{h}_t^*) \) denote the voluntary test participation function. Individual test participation is

\[
p_t = \begin{cases} 
\varrho_z^v(s_{l-1}, t, h_t^*, \overline{h}_t^*) & \text{if } \theta \in \mathcal{P}(\{t_0, t_1, \ldots, t_J\}) \\
\varrho_z(s_{l-1}, t, h_t^*, \overline{h}_t^*) & \text{if } \theta = \text{NA}.
\end{cases}
\]

The presence of a screening policy itself could affect the patient’s beliefs about the usefulness of screening. Therefore, a non-conforming patient’s test participation pattern in the presence of screening policy is, in general, different from his voluntary test participation pattern in the absence of screening.\(^{20}\)

A special case is that the patient fully conforms to the screening policy. Then, participation in the presence of screening policy takes the following form:

\[
\varrho_z^s(s_{l-1}, t, h_t^*, \overline{h}_t^*; \theta) = \begin{cases} 
1 & \text{if } s_{l-1} = 0 \text{ and } t \in \theta \text{ and } h_t^* \neq h^*_N \text{ and } \overline{h}_t^* \neq \overline{h}_N^* \\
0 & \text{if } t \notin \theta \text{ and } h_t^* \neq h^*_N \text{ and } \overline{h}_t^* \neq \overline{h}_N^* \\
0 & \text{if } h_t^* = h^*_N \text{ or } \overline{h}_t^* = \overline{h}_N^*
\end{cases}
\]

where \( h_t^* = h^*_N \) means the patient dies due to the disease of interest; \( \overline{h}_t^* = \overline{h}_N^* \) means the patient dies due to any disease other than the disease of interest.

\(^{19}\)It is worth noting that a recommendation against screening in all periods (i.e., \( \emptyset \) in the power set) is different from no screening recommendation in all periods (i.e., \( \text{NA} \)).

\(^{20}\)In a more general case, a screening policy could include reimbursement for testing. There, the need to differentiate between non-conforming participation and voluntary participation is stronger because one’s budget constraint would be altered in the presence of the screening policy.
2.3.3 Health state transition probabilities

We model the process by which latent health evolution is informed by screening results using health state transition probabilities. The following considerations motivate our specification of policy-relevant health transitions: current diagnosis correctness and historical diagnosis correctness. As previously mentioned, the $n$ test results can be categorized as true positive, true negative, false negative, and false positive given the underlying true health state. In terms of current (period $t$) treatment for disease (and hence health transitions from period $t$ to $t + 1$), a true positive result on a test in time $t$ enables a patient to benefit from treatment, a true negative result avoids unnecessary treatment, a false negative result defers treatment, and a false positive result leads to unnecessary treatment. Hence, current diagnosis correctness may impact subsequent evolution of one’s true health state.

Additionally, results on previous tests may affect current health evolution through two channels. First, historical treatments initiated by previous positive test results can have lagged effects, meaning that the effect of treatment in a previous period may not be fully captured by the current period diagnosed health state. Second, care providers may improve treatment effectiveness through learning based on the observed health outcomes associated with previous treatments.

Due to these considerations, we adopt a high-order Markovian process to characterize the evolution of policy-relevant health. Specifically, we allow age-specific transitions among policy-relevant health states to vary across diagnosis histories as well as current diagnosis correctness.\footnote{Because we wish to focus on the interplay between health transition and screening for a particular disease, we assume that transitions among policy-irrelevant health states are standard Markovian with no memory.}

We begin by specifying the number of possible diagnoses at any period $t$. We note that each of the $n$ possible latent policy-relevant health states may be diagnosed as any one of the $n$ possible disease states. So, for any single period $t$, there are $n^2$ possible combinations of current policy-relevant health state and current diagnosed disease state (i.e., $n^2$ types of diagnoses that correspond to the row and column labels of the confusion matrix).\footnote{For each period, there is one type of true negative patient, $n - 1$ types of false negative patients, $n - 1$ types of false positive patients, and $n - 1 \times n - 1$ types of true positive patients. In total, there are $[1 + (n - 1) + (n - 1) + (n - 1)^2] = n^2$ types of diagnosis correctness.}
the literature, a patient’s diagnosis is conventionally summarized by a table with two rows and two columns (called the table of confusion) that reports the number of true negative (TN), false positive (FP), false negative (FN), and true positive (TP) outcomes. Following this convention, we specify four age-specific health transition functions based on the period $t$ diagnosis correctness and labeled $H_t^{TN}(\cdot), H_t^{FP}(\cdot), H_t^{FN}(\cdot),$ and $H_t^{TP}(\cdot)$. We use these functions later when we fully characterize the health evolution process.

Having introduced the health transition function notation, we discuss the arguments of these functions with specific attention given to the incorporation of diagnosis history. The probability of a particular latent health state in period $t + 1$ depends on the unobserved health state at period $t$ as well as previous observed diagnoses. Given the number of possible diagnosis outcomes ($n^2$) and, hence, the number of possible diagnosis histories up to period $j \leq J$, namely $(n^2)^j$, we are motivated to simplify the history based on reasonable assumptions about disease diagnosis.

Our first assumption is that a positive disease diagnosis ($s_t^d = 1$) is irreversible. In theory, a patient with the disease of interest may return to a health state free of that disease. In practice, however, it is technically difficult to determine whether a reverse transition of health state actually occurred if the only observable data are a positive disease indicator in period $t$ and a diagnosed health state “free of disease” in some subsequent period $t + j$. Furthermore, disease registry databases (e.g., SEER-Medicare) typically include information about initial disease diagnosis only, and it is unlikely that a surveyed individual receiving a positive disease diagnosis would be reclassified as being negatively diagnosed some time later. Thus, we make the following assumption about the disease indicator.

**Assumption 1.** An individual who receives a positive result on a confirmatory diagnostic test in period $t$ retains that positive disease status for the rest of his or her life. That is, if $s_t^d = 1$, then $s_{t+j}^d = 1$ for all $j = 1, \ldots, J$ or until death.

Based on Assumption 1, we rule out the possibility that a patient with a negative disease indicator (i.e., true negative or false negative) in period $t$ has ever received a positive diagnosis for the disease of interest and, in turn, has received treatment for that disease in

---

23 Consider a disease with only $n = 4$ ordinal states. If a researcher wishes to trace the diagnosis history for, say, $j = 10$ previous periods, he needs to examine $4^{2 \times 10} = 1.1e12$ possible histories, which is computationally intractable.
the past. Therefore, it follows that the age-specific policy-relevant health transition probabilities for an individual with a true negative (TN) or false negative (FN) result (in terms of diagnosis correctness) do not depend on that individual’s history of diagnosis. Put differently, that individual’s history of diagnosis does not vary. According to Equation 3
\[ \Delta(h_t^*; \psi) = [0 \ h_t^d] \forall t, t = t_0, \ldots, t \text{ and for all possible latent health states up to period } t. \]

However, the history of diagnosis plays an important role for a patient with a true positive (TP) or false positive (FP) confirmatory test result. Specifically, the history defines the first period of detection of disease (with the understanding that the positive diagnosis may have been made in error). We care about the first period of detection (or earliness of disease diagnosis) because all treatment post-diagnosis is likely to influence evolution of the latent policy-relevant health. While we propose to consider the role of historical diagnoses on current health, we recognize that effects of past diagnosis (and treatment) likely decay with time. That is, we might expect there to be a limit on the length of time over which lagged treatment effects persist. Hence, we make a second assumption regarding the diagnosis history.

**Assumption 2.** We assume that only the most recent \( t - m \) periods of an individual’s diagnosis history affect current health if disease is suspected (i.e., true positive or false positive confirmatory test).

Assumption 2 requires that a researcher consider only \( (n^2)^{t-m} = n^{2(t-m)} \) different diagnosis histories, which is considerably smaller than keeping track of one’s entire history. Let \( c_t = \{[h_{m}^d, \ldots, h_{t-1}^d], [h_{m}^*, \ldots, h_{t-1}^*]\} \) denote the diagnosis correctness history for \( t - m \) periods prior to period \( t \) for a true positive or false positive patient.

We can now define the transition functions for both policy-relevant health and policy-irrelevant health in the presence of information gained through screening. The notation \( \pi(\cdot) \)

---

24In the empirical work that follows, we found that the number of possible diagnosis histories is exceedingly large even when we restrict the history to the last \( t - m \) periods. In practice we had to place additional restrictions on the relevant history. We summarize the history of diagnosis by the length of time since the first positive presumptive test, \( t - m \), and the diagnosed disease state at that time, \( h_m^d \). Note that the former value defines the “earliness of diagnosis”. In the theoretical model that follows, \( c_t \) would be replaced by these two components, \( (h_m^d, t - m) \), rather than the history since period \( m \) (\( c_t \)).
simply denotes the probability of a period $t + 1$ health outcome. More specifically,

$$\pi(h^*_t) = H_t(h^*_t, c_t, s_t^d) = \begin{cases} 
H^T_{TN}(h^*_t) & \text{if } s_t^d = 0 \text{ and } h^*_t = h^*_1 \\
H^F_{FP}(h^*_t; c_t) & \text{if } s_t^d = 1 \text{ and } h^*_t = h^*_1 \\
H^F_{FN}(h^*_t) & \text{if } s_t^d = 0 \text{ and } h^*_t \in H^* \setminus h^*_1 \\
H^T_{TP}(h^*_t; c_t) & \text{if } s_t^d = 1 \text{ and } h^*_t \in H^* \setminus h^*_1 
\end{cases}$$

(6)

$$\pi(h^*_{t+1}) = H_{t+1}(h^*_{t+1})$$

(7)

where the diagnostic health outcomes are updated as follows:

$$s_t^d = \begin{cases} 
s_{t-1}^d, & \text{if } p_t = 0 \\
[\Delta(h^*_t; \psi)]_1 & \text{if } p_t = 1 
\end{cases}$$

$$h_t^d = [\Delta(h^*_t; \psi)]_2$$

(8)

and the length of one’s relevant diagnosis history is $t - m$, where

$$m = \begin{cases} 
\min\{j | s_j^d = 1\} & \text{if } \sum_{j=t_0}^{t} s_j^d \geq 1 \\
\text{undefined} & \text{if } \sum_{j=t_0}^{t} s_j^d = 0 
\end{cases}$$

(9)

For completeness we also specify that the health transition process is terminated upon diagnosed death associated with either process. That is,

$$h^*_t = h^*_n, \text{ if } h_{t-1}^* = h^*_n \text{ or } \bar{h}_{t-1}^* = \bar{h}_n^*$$

and

$$\bar{h}_t^* = \bar{h}_n^*, \text{ if } h_{t-1} = h^*_n \text{ or } \bar{h}_{t-1} = \bar{h}_n$$

(10)

Recall that the model of latent health transition has three main components: the policy-specific screening participation function (equation 5); the diagnostic test function (equation 3); and the latent health transition functions (equations 6 and 7).

### 2.4 Reduced-form representation

Let $h^*_t$ be an indicator vector of length $n$: if the patient resides in the $k^{th}$ policy-relevant health state, then $[h^*_k] = 1$ and the other $n - 1$ elements are 0. Similarly, we define $\bar{h}_t^*$ for policy-irrelevant health states. Let matrix $O_h$ denote the sequence of health states from time period $t_1$ to $t_J$. Specifically,

$$O_h = \begin{bmatrix} 
h_{t_1}^* & h_{t_2}^* & \ldots & h_{t_J}^* \\
\bar{h}_{t_1}^* & \bar{h}_{t_2}^* & \ldots & \bar{h}_{t_J}^* 
\end{bmatrix}_{(n+\pi) \times J}$$

(11)
Using this notation, we can rewrite the structural health evolution process into its reduced-form:

$$O_h = R_z(h^*_t, s^d_t, \overline{h}^*_t, \theta) = \left( [\Delta^p(\cdot; \psi^p) \Delta^c(\cdot; \psi^c)]^T \circ [\varrho^p(\cdot) \varrho^c(\cdot)]^T \circ \begin{bmatrix} H_t(\cdot) & \overline{H}_t(\cdot) \end{bmatrix}^T \right)(h^*_t, s^d_t, \overline{h}^*_t, \theta).$$  (12)

Four factors govern $R_z(\cdot)$. They are: (1) the accuracy of presumptive and confirmatory tests ($\psi^p$ and $\psi^c$), (2) the patient participation function in the presence of screening ($\varrho^p(\cdot)$) and the participation function in the absence of screening ($\varrho^c(\cdot)$), (3) the treatment effectiveness in the policy-relevant dimension ($H_t(\cdot)$), and (4) the background health transition in the policy-irrelevant dimension ($\overline{H}_t(\cdot)$). All of these are exogenous factors that are statistically independent from any particular screening policy ($\theta$). Thus, the functional form of individual health evolution is screening policy invariant.

### 2.5 Optimal screening policy in a principal-agent problem

The policymaker chooses a screening policy at the beginning of period $t_0$ to maximize the expected social benefit collected over the study time horizon ($t_0, \ldots, t_J$) from a population. The representative individual’s initial policy-relevant health state ($h^*_t$) and policy-irrelevant health states ($\overline{h}^*_t$) are drawn from a conditional (upon $z$) joint distribution $G(\cdot)$. Specifically,

$$[h^*_t, \overline{h}^*_t]^T \sim G([h^*_t, \overline{h}^*_t]^T | z).$$  (13)

The social payoff of a particular health state is defined, here, as the monetary value of quality-adjusted life years (QALY) net of the sum of diagnostic test costs and care costs under the assumption that there is a uniform social willingness to pay for a QALY. Let $\alpha = [\alpha_1 \ldots \alpha_n]^T$ and $\overline{\alpha} = [\overline{\alpha}_1 \ldots \overline{\alpha}_n]^T$ denote the QALY adjustment factor vector for the $n$ policy-relevant health states and the $n$ policy-irrelevant health states, respectively. Let $\gamma$ denote the time discount factor. According to the comparative effectiveness criterion, the optimal screening policy function, denoted $\theta^*(\cdot)$, is

$$\theta^* \left( [\psi^p, \psi^c]^T, [\varrho^p(\cdot) \varrho^c(\cdot)]^T, [\alpha, \overline{\alpha}]^T, [H_t(\cdot) \overline{H}_t(\cdot)]^T, G(\cdot | z), F(z), \gamma \right)$$  (14)

$$= \arg \max_{\theta \in \Theta} \int \int [\alpha, \overline{\alpha}]^T \cdot \sum_{t=t_1}^{t_J} E_{t_0} \left[ [O_h]_{t,t} \right] \gamma^{t-t_0} dG([h^*_t, \overline{h}^*_t]^T | z)dF(z)$$
where \( E_{t_0} [\cdot] \) is the expectation at the beginning of the study time horizon and \( F(z) \) is the distribution of time-invariant characteristics \( z \) in the study population\(^{25} \). We further characterize social benefit frontier in Appendix 1.

### 2.6 Empirical estimation strategy

Clearly, one’s true health at each period of his life is unobserved. Hence, we cannot use such (non-existent) data to directly estimate the parameters that define these true health transitions (equations 6 and 7). Let us consider, however, two alternative approaches that would determine the modeled health transition functions.

The first approach is to use the model and an initial set of parameters to predict latent health outcomes each period and update the disease indicator and diagnosed disease state in each period. We would then use a simulated maximum likelihood procedure or simulated method of moments procedure to match the observed diagnostic test results with the simulated diagnostic test results produced by the model. Unfortunately, generation of a diagnostic health history for each possible series of latent policy-relevant health is analytically intractable. An alternative approach recognizes that the health transition functions (equation 6) can be estimated if one’s true (policy-relevant and policy-irrelevant) health is observed. Hence, we define three conditions that, if satisfied, would be necessary and sufficient for estimation of the health transition functions of interest. The conditions are

- **Condition 1**: for a diagnosed patient, the researcher knows the patient’s policy-irrelevant health state at any post-diagnosis period.
- **Condition 2**: a diagnosed patient’s diagnosed disease state is correct.
- **Condition 3**: for a diagnosed patient, the researcher knows the patient’s policy-relevant health state at any post-diagnosis period.

First, we note, Condition 1 is not demanding. Given our expressed desire to measure the effect of screening on policy-relevant health, we need only summarize the policy-irrelevant health state with a binary indicator of death. By assumption, it is only when death in the

\(^{25}\)It is worth noting that if life expectancy is the outcome of interest, then by setting (1) all the QALY adjustment factors across non-death states to an arbitrary positive constant, (2) the QALY adjustment factors for death (\( \alpha_n \) and \( \pi_n \)) to 0, and (3) time discounting factor (i.e., \( \gamma \)) to 1, equation 14 still holds.
policy-irrelevant dimension occurs that health evolution in the policy-relevant dimension is affected. The following two assumptions about a confirmatory test would lead to the satisfaction of Condition 2. They are

**Assumption 3.** A confirmatory test is perfectly specific.

**Assumption 4.** A confirmatory test correctly ascertains the disease state for a true positive patient.

Assumption 3 states that all individuals who are free of disease are correctly diagnosed. Under Assumption 4, the diagnosed health state of a diseased individual is the true health state. It is worth noting that these two assumptions are usually considered to be true in empirical work. Condition 3 can be met if we consider other patient-level information available in most medical databases. Admittedly, a patient may not receive a confirmatory diagnostic test measuring policy-relevant health each period, which under Assumption 4 would provide us the latent health state each period. However, upon being positively diagnosed, a patient is likely to undergo treatment over time. Medical care consumption is available in claims data. Medical treatment patterns often reflect doctor-suspected health changes. Hence, a patient’s post-diagnosis treatment utilization history obtained from medical claims data goes a long way toward satisfying Conditional 3.

Under these conditions, latent health is observed and the health transition functions proposed above can be estimated directly. Note that the health transition functions are screening-policy independent as shown before. Having recovered the parameters of the health transition functions, a researcher simply needs to specify a new screening policy in order to evaluate changes in recommended ages of diagnostic testing, or specify a new confusion matrix for evaluating the impact of a new presumptive or confirmatory diagnostic test.

---

26 We recognize that death transitions within the policy-relevant dimension may depend on health in general, but we abstract from that consideration here. However, it would not be difficult to consider other measures of health, such as limitations with activities with daily living (ADLs), as components of patient’s exogenous characteristics, \( z \in \mathbb{R}^K \).

27 A proof of satisfaction of condition 2 is shown in Appendix 2.
3 Application Example: PSA Screening

This section provides an application of our model. Specifically, we use our model to assess the impact of recommended age of screening and screening test accuracy on life expectancy among the 20 million new male Medicare beneficiaries with no prostate cancer diagnosis history prior to age 65. As explained before, a reduced-form analysis that uses the screening intervention as the explanatory variable of interest would require prospectively following patients over their post-screening lives in order to predict screening’s life-cycle impact. It is faster and less expensive to collect medical care claims data that retrospectively track patient history. We seek to demonstrate our model’s ability to utilize claims in predicting the lifetime effects of a wide range of screening policies. Therefore, we chose the 1992-2006 SEER-Medicare database as our analysis data. The SEER-Medicare database consists of linked information from two large population-based sources of data. The Surveillance, Epidemiology and End Results (SEER) program details clinical, demographic, and cause of death information for persons with cancer. Information on medical care utilization of covered services is available from Medicare claims, which are recorded from eligibility (usually at age 65) until death.

3.1 Empirical model of health transitions under PSA screening

Here, we refine the general model with the available data in mind. Specifically, the empirical model explains the health transitions of males from age $t_0 = 65$ (when one becomes eligible for Medicare) until the age of 100. We assume that individuals have no prior diagnosis of prostate cancer (PCa). At any period $t$, the set of possible (true) health states includes: no PCa, localized-PCa, advanced-PCa, and all-cause death. No PCa corresponds to a health state free of the policy-relevant disease. Localized- and advanced- PCa correspond to two characterizations, or subtypes, of disease progression. We do not differentiate between death caused by the policy-relevant disease or by another cause for two reasons. Clearly, death in the policy-irrelevant dimension is important because it stops the policy-relevant health transition process; hence we must account for the probability of such an event. Also, the

\[^{28}\text{US Census Bureau life tables suggest that life expectancy beyond age 100, given an age of 100, is very small (US Census Bureau 2007).}\]
The diffusion of PSA testing over the last three decades has led to an increased attribution of death to prostate cancer (Feuer et al., 1999). Such attribution bias may overstate the probability of PCa death in clinically-observed data.\footnote{We explore possible model misspecification that arises from this choice of all-cause death in more detail later.}

Under the assumption of no reversal of disease diagnosis, and the assumption that an advanced disease state would not be re-classified as a localized disease state in most available patient-level databases, there are nine possible true health state transitions. Additionally, the confirmatory test for prostate cancer (i.e., biopsy) has a nearly perfect specificity (Haas et al., 2007). Thus, our empirical model does not require specification of a health transition function for false positive diagnoses. The empirical health transition specification becomes

\[
\pi(h_{t+1}^*) = \begin{cases} 
H_{t}^{TN}(h_t^*) & \text{if } s_t^d = 0 \text{ and } h_t^* = h_1^* \\
H_{t}^{FN}(h_t^*) & \text{if } s_t^d = 0 \text{ and } h_t^* \in H^* \setminus h_1^* \\
H_{t}^{TP}(h_t^*, h_m^d, t - m) & \text{if } s_t^d = 1 \text{ and } h_t^* \in H^* \setminus h_1^* 
\end{cases}
\]  

(15)

where the diagnostic health outcomes are updated according to equation 8 and the length of one’s relevant diagnosis history is characterized by equation 9 for \( t = 65, 66, \ldots, 100 \). We also specify an initial policy-relevant health state distribution function to describe the probabilities of different latent health states at \( t_0 = 65 \).

### 3.2 Estimation using medical claims data

While there are only four possible health states in any period \( t \) of our empirical model, there are \( 35 - t \) possible lengths of life (based on years) for each possible age \( t > t_0 \). That is, an individual observed at age \( t_0 = 65 \) can live one year, two years, et cetera, up to 35 years (to age 100). Each of those life sequences consists of transitions between the possible health states. Hence, there are nearly 5,000 transition probabilities that govern the health evolution process defined in Equation [15]. Predictions of those transitions depend on the length of time since disease diagnosis and other individual characteristics. Despite the richness and magnitude of claims data samples (as opposed to the small sample trial data), we do not have enough data to estimate the parameters that differentiate these theoretically-possible transitions. For example, because the SEER-Medicare data contain records of those who have been diagnosed with cancer, we cannot estimate the probability of transiting from a health
state free of PCa to one of localized- or advanced-PCa. In light of these estimation challenges, we explain in detail below our strategies for estimating health transitions associated with true negative, true positive, and false negative screening outcomes.

[Figure 1 about here.]

3.2.1 True negative health transition

As shown in Figure 1, the subsequent (i.e., next period) health state of an individual who has tested negative for PCa and whose true health state is PCa-free is to remain free of the policy-relevant disease, or to transition to either localized-PCa or death by any cause. We assume that natural disease progression takes some time and that the diagnostic test would detect PCa if it was present. Therefore, a person who correctly tests negative for disease, cannot transition to an advanced stage of the disease in one period. Since the SEER-Medicare database contains only those individuals ever diagnosed with cancer, we cannot use these data to measure cancer incidence. Instead, we assign the probability of transition from PCa-free to localized-PCa (2.33%) based on published data in the Prostate Cancer Prevention Trial (PCPT).30

We obtain the mortality rate at each post-65 age from the 1992-2006 SEER-Medicare five percent PCa-free random Medicare beneficiaries sample that accompanies the cancer database. This specific sample (n=39,297) is constructed from persons who have no cancer and persons who have cancers other than PCa with an effort to match the complement of the SEER sample (i.e., the prostate cancer incidence sample). We impose a mortality rate at age 100 of 100%. We solve for the probability of maintaining a PCa-free state as one minus the sum of the progression to localized-PCa and all-cause morality rates at each age. Figure 2 depicts the age-specific mortality rate for true negative patients recovered from SEER-Medicare PCa-free matching sample. As expected, the mortality rate increases as age increases.

[Figure 2 about here.]

30See Appendix 3 for an explanation of estimation of this probability.
True positive health transition

We are able to use the observed data of the SEER-Medicare cancer patients to estimate true positive health transitions. The SEER-Medicare database links Medicare-eligible persons with cancer in the SEER data with their Medicare claims. The resulting files provide a unique population-based source of information about health and medical care use that spans a continuum of care from the period of initial diagnosis and treatment to long-term follow-up and care. Specifically, the SEER-Medicare database records the age at prostate cancer diagnosis ($m$), the diagnosed prostate cancer disease state ($h_{m}^{d}$) at diagnosis, and, if death occurs during follow-up, the death age for Medicare beneficiary prostate cancer incidents.

One’s age-specific diagnosis duration ($t - m$) and his initial diagnosed disease state ($h_{m}^{d}$) defines the earliness of diagnosed disease. In order to estimate the true positive health transitions, $H_{t}^{TP}(h_{t}^{*}; h_{m}^{d}, t - m)$, we need to identify all possible health state transitions of those with the disease. A patient who was initially diagnosed with advanced-PCa may either remain in this health state or die. Because death is observed in the SEER-Medicare data, we can estimate these two health transition probabilities.

The challenge is to estimate health state transition probabilities for patients who were initially diagnosed with localized-PCa. For these patients, we need to know their health state at any post-diagnosis time in order to measure the three possible health state transitions. In particular, we need evidence of disease progression from localized-PCa to advanced-PCa in the SEER-Medicare database. As mentioned above, however, most data sets record the initial diagnosed disease state but do not update the disease state. We use the available medical care claims data to fill in this gap. That is, among patients initially diagnosed with localized-PCa who used surgery (i.e., prostatectomy) as their initial treatment, we consider the first subsequent use of androgen deprivation therapy as an indicator of the onset of advanced-PCa.

We use patients who were diagnosed with advanced-PCa between 1992 and 2005 ($n = 15824$) for similarly diagnosed patients who use radiation therapy as an initial treatment, androgen deprivation therapy cannot be used to signal advanced-PCa onset because it is often used as an auxiliary treatment in conjunction with the main initial treatment.

We restrict the sample ($n = 19483$) to patients who (1) were diagnosed with localized prostate cancer between 1992 and 2005; (2) have only Part A and Part B Medicare plans over the entire post-diagnosis follow up; and (3) underwent surgery as their initial treatments.

---

31 We use patients who were diagnosed with advanced-PCa between 1992 and 2005 ($n = 15824$).
32 For similarly diagnosed patients who use radiation therapy as an initial treatment, androgen deprivation therapy cannot be used to signal advanced-PCa onset because it is often used as an auxiliary treatment in conjunction with the main initial treatment.
33 We restrict the sample ($n = 19483$) to patients who (1) were diagnosed with localized prostate cancer between 1992 and 2005; (2) have only Part A and Part B Medicare plans over the entire post-diagnosis follow up; and (3) underwent surgery as their initial treatments.
It remains, then, to quantify one of the main determinants of these health state transitions: diagnosis duration. Because duration of disease is continuous, we round (down) diagnosis duration into one-year integers. Thus, a patient of age $t$ has $t - 65 + 1$ possible diagnosis durations. For example, a 70-year old patient in our sample could have been living with his diagnosed disease for 6 ($= 70 - 65 + 1$) possible durations: 0 years, 1 year, $\ldots$, 5 years. As one might expect, we do not observe lengthy diagnosis durations in the SEER-Medicare data. Despite the large age range in these data, observations of a 90-year old with PCa who was initially diagnosed twenty years prior are rare. To address this frequency issue, we first examine age and diagnosis duration combinations in the SEER-Medicare 1992-2006 database to ascertain the duration distribution and maximum duration across all ages. We found sufficient data to track health transitions at all ages for those with up to eight years of diagnosed PCa. This data limitation (i.e., too few observed age-specific health transitions among those with greater than eight years duration) forces us to treat health transitions that depend on combinations of age and diagnosis data that result in greater than eight years duration the same as those with exactly eight years duration. Once these data are used to estimate transition probabilities as a function of duration, we can make predictions of health behavior beyond eight years through extrapolation.

Under our assumptions stated in this section and given the initial diagnosis disease state and a diagnosis duration, we can estimate the 4,662 age-specific transition probabilities for true positive patients beginning at age 65.\footnote{Assumption 2 above makes this limitation inconsequential, as we will see later.} Figure $\text{3}$ shows the age-specific mortality rates of true positives who have localized-PCa and advanced-PCa, respectively. In the figure, we present transitions for three diagnosis duration categories: duration equal to zero, equal to two, and greater than five years for localized-PCa patients and duration less than two, between two and five, and greater than or equal to five years for those with advanced-PCa. Localized-PCa patients have much lower mortality than advanced-PCa patients across all ages. Among true positive patients with localized-PCa currently, earlier diagnosis (indicated

\footnote{Given an age and diagnosis duration, the true positive patients with localized-PCa can transition to three health states. Patients who have advanced-PCa and were originally diagnosed with either localized- or advanced-PCa can transition to two possible health states. In addition, as explained before, at age $t$, there are $t - 65 + 1$ possible diagnosis durations. Hence, there are $4,662 (= (3 + 2 + 2) \sum_{t=65}^{100} (t - 65 + 1))$ transition probabilities (not including the probability of 1 for transition from death to death) for true positive patients.}

23
by a longer diagnosis duration) has little impact on mortality rates. This empirical information reflects the general opinion that localized-PCa is not a deadly disease and suggests that earlier treatment (or diagnosis) does not improve survival significantly. In contrast, among true positive patients with advanced-PCa currently, those with a longer diagnosis duration are less likely to die. This empirical evidence reflects the opined benefit of early diagnosis.

[Figure 3 about here.]

Figure 4 shows the age-specific probability of progression from localized-PCa to advanced-PCa across diagnosis durations for true positive patients. Among such patients, those with a diagnoses duration less than one year are more likely to progress to advanced-PCa than those with a longer diagnosis duration. Among patients with a diagnosis duration longer than one year, there is little variation by duration of the age-specific probabilities of progression to advanced-PCa. This finding suggests that early diagnosis of localized-PCa helps slow disease progression, yet the slowing effect diminishes rapidly as diagnosis duration increases.

[Figure 4 about here.]

Our prediction of transition probabilities for health state combinations with diagnosis duration longer than eight years is based on extrapolation. Because these probabilities are bounded from below,\textsuperscript{36} the more rapidly the marginal effect of duration (i.e., historic diagnosis correctness) diminishes between durations of zero years and eight years, the more accurate is the extrapolation assumption for diagnosis durations greater than eight years.

The duration-specific probabilities of transition from localized-PCa to advanced-PCa among patients who are currently diagnosed with localized-PCa are provided in Figure 5. Patients are grouped into three age categories: between 65 and 74, between 75 and 84, and age 85 and above. The figure depicts the rapidly diminishing marginal effect of diagnosis duration on health transition (i.e., the reduced probability of disease progression) across all age groups. Transition probabilities vary little (by duration) beyond two years post-diagnosis.

[Figure 5 about here.]

\textsuperscript{36}We maintain that they are bounded from below because disease progression is most likely inevitable, and at a minimum, zero.
Figure 6 depicts how the probability of transiting from advanced-PCa to death changes as diagnosis duration increases among true positive patients with advanced disease currently who were also initially diagnosed with localized-PCa. As in Figure 5, the marginal effect of diagnosis duration on health transition diminishes rapidly (over five years) across age categories.

[Figure 6 about here.]

Figure 7 shows how the probability of transiting from advanced-PCa to death changes as diagnosis duration increases among true positive patients with advanced-PCa currently and who were initially diagnosed with localized-PCa. The transition probabilities are relatively constant as diagnosis duration increases for each age category. That is, among patients who were first diagnosed with localized-PCa, earliness of diagnosis has little impact on mortality, conditional on currently being in an advanced-PCa state. Together, Figures 5, 6, and 7 support our truncation of the observed diagnosis durations to eight years in estimation and the use of extrapolation in our predictions of transition beyond eight years.

[Figure 7 about here.]

### 3.2.2 False negative health transition

Because it is unethical to knowingly misinform a diseased patient with a negative diagnosis, a false negative sample does not exist in the observed data. To circumvent this limitation, we assume that the health state transition probabilities from $t$ to $t + 1$ of someone who has PCa would not differ if he was newly diagnosed at $t$ or remained undiagnosed at $t$. Our assumption is based on evidence that it takes more than one year for post-diagnosis treatments to manifest their curative (or harmful) effects (Bill-Axelson et al., 2005, 2008). We recognize that this assumption is open to clinical debate. Based on the assumption, we can use the health transition functions for the true positive patients diagnosed within one year to approximate those of false negative patients. As indicated in Figure 1, a localized-PCa patient can transition to three possible states (i.e., localized-PCa, advanced-PCa, or death) and an advanced-PCa patient can transition to two possible states (i.e., advanced-PCa or death). Therefore, over a lifetime beginning at age 65, $108 \ (= (100 - 65 + 1) \times 3)$
probabilities and 72 \((= (100 - 65 + 1) \times 2)\) probabilities define the health transitions for a false negative patient with localized-PCa and a false negative patient with advanced-PCa, respectively.

4 Policy Experiments

We use the estimated model of health evolution informed by screening to evaluate alternative screening policies. Specifically, we examine how age-specific recommendations for prostate screening affect life expectancy. The U.S. Food and Drug Administration has approved a screening test to detect prostate cancer in men age 50 and older based on prostate-specific antigen (PSA), which is a protein produced by cells of the prostate gland. Levels of PSA are measured in nanograms per milliliter (ng/mL) of blood. Higher levels of PSA indicate a greater likelihood that cancer is present. A PSA level below 4.0 ng/ml is generally considered normal, but because of variation in PSA levels due to non-cancer related reasons, the accuracy of the PSA test is reported for reference ranges, or cut-off levels. If prostate cancer is suspected, a biopsy is necessary to determine whether cancer is present. Table 4 displays the sensitivity (i.e., true positive rate) and specificity (i.e., true negative rate) accuracies of the PSA test for four different cut-off values. The sensitivity of a prostate cancer biopsy is also reported. Sensitivity is reported for both localized and advanced disease.

Table 1: Sensitivity and Specificity of PSA Screening Tests and Biopsy

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Localized-PCa</td>
<td>Advanced-PCa</td>
</tr>
<tr>
<td>PSA Screening Test(^a)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cut-off: 1 ng/ml</td>
<td>0.856</td>
<td>0.960</td>
</tr>
<tr>
<td>cut-off: 2 ng/ml</td>
<td>0.557</td>
<td>0.860</td>
</tr>
<tr>
<td>cut-off: 3 ng/ml</td>
<td>0.339</td>
<td>0.710</td>
</tr>
<tr>
<td>cut-off: 4 ng/ml</td>
<td>0.209</td>
<td>0.531</td>
</tr>
<tr>
<td>12-core Needle Biopsy</td>
<td>0.60</td>
<td>0.95</td>
</tr>
</tbody>
</table>

\(\text{a} \) PSA test sensitivity and specificity are obtained based on Thompson et al. (2005).

\(\text{b} \) Source: Haas et al. (2007)
Using our estimated model of health transitions, we calculate life expectancy based on an annual PSA screening program that specifies the age at which annual screens would cease. We present results for different recommended “last ages” for PSA tests with four different cut-off levels (i.e., 1 ng/ml, 2 ng/ml, 3 ng/ml, and 4 ng/ml). We predict the effects of these policies on life expectancy through Monte Carlo simulation. For each policy (i.e., last age recommendation and PSA cut-off level), we simulated 500,000 hypothetical Medicare beneficiaries to progress through our estimated model of the health evolution process. Before discussing the results, we mention a few details necessary for simulation.

4.1 Initial conditions

To conduct the policy experiments, we need to specify an initial health state distribution at age 65; we do not want to assume that everyone is initially disease-free. Based on a review of the literature, about 22% of individuals have prostate cancer at age 65 (Coley et al., 1997). In the SEER-Medicare 1992-2006 sample, we found that 4.633% of diagnosed patients have advanced-PCA. Hence, we calculate an initial prevalence rate for localized-PCA and advanced-PCA of 20.98% (= 22% × (1−0.04633)) and 1.02% (= 22% × 0.04633), respectively.

In the policy simulations, we use this initial health state distribution to define the true initial health state. We assume, however, that individuals in the simulation have no history of disease diagnosis. That is, prior to age 65, they have not been screened for prostate cancer.

4.2 Voluntary use

In the absence of a PSA screening policy, Medicare beneficiaries may be screened voluntarily. In light of this behavior, the measured effect of a screening policy should be the difference between the life expectancy under the policy and that in a voluntary use scenario. To estimate the latter baseline life expectancy, we use data on PSA voluntary use from the PLCO study (Andriole et al., 2009). The study reports two types of PSA test users. The
first type is the persistent user who would take a PSA test annually regardless of the PSA screening recommendation. The second type is a non-persistent user who would take a PSA test sporadically. They estimate that the elderly population consists of 9.8% persistent PSA test users and 90.2% non-persistent users. Among the non-persistent users, the probability of taking a PSA test in any one year is 0.1457⁹⁷.

Using these statistics, we simulate life expectancy under voluntary use of PSA in two steps. First, we simulate the age-specific health transitions of persistent users and non-persistent users assuming participation in screening at the rates implied by their type. Next, we weight their simulated life expectancies based on their representation in the population (i.e., 9.8% and 90.2%).

### 4.3 Results of alternative policy simulations

Figure 8 reports the life expectancy of an average 65 year-old Medicare beneficiary under screening policies that specify the last age of annual screening. The curved lines correspond to the four different PSA test cut-off values: 1 ng/ml, 2 ng/ml, 3 ng/ml, and 4 ng/ml. The dotted horizontal lines indicate life expectancy under the voluntary use scenarios (with the four different PSA test cut-off values). Obviously, these comparison scenarios do not vary with the last age of recommended screening because that constraint is not imposed in simulation.

The graph illustrates three statistically significant findings. First, higher PSA test sensitivity (i.e., smaller PSA cut-off value) increases life expectancy. Second, given a PSA test sensitivity, life expectancy increases as the last screening age increases. Third, the marginal return of an extra screening year diminishes greatly if the last age of screening

---

⁹⁷ According to Andriole et al. (2009), among the non-persistent users, 31.9% would use PSA testing once in three years. If we define the probability of taking a PSA test in any given year as \( p \), then the percentage of patients who take only one PSA test over three years (assuming no one dies) is \( p(1-p)(1-p) + (1-p)p(1-p) + (1-p)(1-p)p = 3p(1-p)^2 \). Solving the equation \( 3p(1-p)^2 = 0.319 \), we obtain \( p = 0.1457 \) for non-persistent users.
is extended beyond age 75. This last finding suggests that prostate treatment attributable to screening at older ages does not lead to much improvement in survival (even if one is diagnosed with prostate cancer early). We conjecture that this slowing of longevity gains is driven by the high rates of mortality associated with other diseases beyond age 75.

Relative to voluntary use of PSA testing, compliance with an annual PSA screening policy up to a specified age would extend life expectancy by a maximum of 0.491(0.002), 0.573(0.002), 0.501(0.003), and 0.439(0.003) years using a PSA test with cutoff values of 1 ng/ml, 2 ng/ml, 3 ng/ml and 4 ng/ml, respectively.\footnote{Standard errors are provided in parentheses. The standard errors of life expectancies are calculated using a sample size of 20.0334 million, which is the expected number of new male Medicare beneficiaries with no PCa diagnosis history prior to age 65 between 2011 and 2030.}

Compared with screening patients for one single year at age 65, screening patients up to age 100, would increase life expectancy by 0.5908, 0.6198, 0.6129, 0.5517 years for cutoff value of 1 ng/ml, 2 ng/ml, 3 ng/ml and 4 ng/ml, respectively. If the last screening age was 80 (instead of 100), we find that 94%, 95%, 93%, and 95% of these life expectancy gains would be achieved for each of the cut-off values, respectively. This finding is further evidence that PSA screening among men older than age 80 provides little gain in life expectancy.

We remind the reader that we estimated probabilities of progressing from localized-PCa to advanced-PCa based on localized-PCa patients who underwent surgical treatment. Additionally, we apply these estimated probabilities to all true positive patients with localized-PCa. The current medical literature suggests that surgery may have better survival outcomes than other treatments \textit{(Merglen et al., 2007; Bill-Axelson et al., 2011)}. Hence, our model may overstate the screening benefit for survival; these simulations represent upper bounds on PSA screening effects.

5 Discussion and Conclusion

Our model extends the seminal PM model in important ways. Our model covers behavior and outcomes over multiple periods rather than simply a single period. Instead of assuming that
patients follow a policy prescription (chosen optimally by a policymaker), our framework specifically models the decisions of the patient (in terms of policy compliance) and the policymaker. This latter distinction opens the gate for modeling patient non-conformity. With regard to health evolution, our model characterizes age-specific health transitions as functions of the history of diagnosis. Specifically, the transitions depend on the first period of detection of disease and the level of diagnosed disease. The estimated health evolution process does not depend on specific recommendations associated with screening of the policy-relevant disease. That is, the process depends on the potential components of a screening policy, but not on any observed policy in place at the time of data collection (as would be the case when data are obtained through a screening trial). For this reason, we can use our model to predict the effects of a wide range of screening policies, including those that do not currently exist in practice.\textsuperscript{39}

The empirical estimation strategy we present in this paper has two features. First, variation in screening policy (e.g., specified age for first screen, specified duration between screens, specified last age of screening, etc.) is not required for estimation of our model. Our model depends only on the policy components of screening, and not the actual policy in place at the time of data collection. Hence, we require only variation in age, earliness of diagnosis, and the level of initially-diagnosed health. It is worth noting that variation in age is abundant in observational data and can be achieved with relatively low cost (compared with lifetime follow-up) in experiments. In addition, a diagnostic test (presumptive or confirmatory, or their combination), as a stochastic mapping, would also generate \textit{exogenous} variation in both earliness of diagnosis and initially-diagnosed health even if there is no variation in screening recommendation. Therefore, variations required by our empirical estimation approach are not demanding.

Second, the model does not rely on a lengthy follow-up of survey participants. The available data on a large group of individuals covering a large age span allow us to estimate

\textsuperscript{39} Since our health evolution process depends on conditional probabilities (incidence), the model does not suffer from lead time bias. In addition, one can empirically remove length time bias by ensuring that a health state specified in empirical estimation is truly homogeneous from a clinical standpoint.
non-linear age-specific health transition functions. The estimation results provide evidence that the marginal effect of an early diagnosis is diminishing over time, given the initial diagnosed disease state. That is, while the diseased state at the time of detection is statistically relevant, the length of time post-diagnosis does not provide any additional impact on health transitions after two to five years. While health still evolves, it is guided by age, period-\(t\) health, and policy-irrelevant death.

Although our model does not require data from a screening study, it suggests opportunities for improved design of screening policy experiments. Because we found that at most eight years of health state histories for prostate cancer patients at each age were needed to recover the health transition functions, this implies that it is unnecessary for a PSA screening experiment to follow patients over their entire life cycles. Such a study only requires that the age distribution of any incoming study cohort ensures that enough living patients populate each discrete age during an eight-year follow-up. This requirement reveals a trade-off between two cost variables: a larger sample or length of follow-up period. However, an increased sample size and reduced duration of experiment has the added benefit of reducing the time needed for screening policy decision making from decades to, potentially, a couple of years.

Despite its advantages, both estimation of our model and simulated results from our model should be interpreted in the context of several limitations. With regard to estimation, the data limitation of not observing false negative patients requires that we assume that their disease progression during the first period of diagnosis is the same as that for true positive patients in the PSA screening example. Given that it may take time to observe the effects of treatment, this assumption is reasonable. However, the validity of this assumption depends on the length of the modeled health transition: the shorter the time interval the more likely the assumption’s validity.

Additionally, clinical trials have shown that, when compared with watchful waiting, even the best treatment does not slow disease progression for those with localized disease during the first year of treatment (Bill-Axelson et al., 2005, 2008, 2011). For advanced prostate cancer treatment, the standard treatment (i.e., androgen deprivation) is palliative treatment and does not improve survival (Miyamoto et al., 2004).
Although our theoretical model accommodates health evolution in both the policy-relevant and the policy-irrelevant health dimensions, the presence of death-cause attribution bias makes it difficult for model estimation to follow the theoretical model exactly. To avoid this bias, our empirical model applied to PCa transitions and PSA screening relies on all-cause death only. We recognize that we may introduce selection bias if patients who are healthier in the policy-irrelevant dimension are more likely to be screened. We discuss the identification concern due to non-observability of the policy-irrelevant health state in Appendix 4.

When we simulate life expectancy under different screening policies, we assume that individuals adhere to the policy recommendations. This assumption is likely to hold when a screening test is inexpensive or involves little out-of-pocket expense. Screening tests for many diseases are fully reimbursed by insurance, so the full-conformity assumption of our simulations is not very limiting for many applications. It is worth noting that our model can be used to assess upper-bound screening effects when a screening test involves significant out-of-pocket expense.

We conclude by re-emphasizing the ability of our model to evaluate alternative screening recommendations without conducting costly screening interventions. We present a generic theoretical model of health transition that can translate observable disease characteristics (i.e., disease states following a natural history along with a probability mass function defining initial disease states), diagnostic test accuracy, and screening policy (i.e., timing of screening) into survival outcomes. This model can be used for comparative effectiveness assessment of alternative screening test utilization policies over multiple periods. In addition, our model provides the groundwork for determining optimal screening policy over multiple time periods inside a single expected pay-off maximization problem in the spirit of Phelps and Mushlin [1988]. Our model’s health transition functions capture treatment effectiveness and do not depend on a particular screening recommendation. Rather, they depend on the components of any possible screening policy. Screening for disease will alter the earliness of disease diagnosis. Our model depends on this earliness of detection, as well as age and
diagnosed health levels. Simulations from our estimated model applied to prostate cancer suggest that earliness of disease diagnosis has a diminishing marginal effect on health transitions after a short period of time. Because estimation of our model of health transitions is achievable using medical care claims and retrospective data, we avoid the potentially lengthy time requirements of data collection based on screening interventions.

References


URL: http://www.cancer.org/Healthy/FindCancerEarly/CancerScreeningGuidelines


URL: http://www.auanet.org/content/press/press_releases/article.cfm?articleNo=107


**URL:** [http://www.uspreventiveservicestaskforce.org/uspstf/uspsprca.htm](http://www.uspreventiveservicestaskforce.org/uspstf/uspsprca.htm)

URL: http://www.uspreventiveservicestaskforce.org/uspstf/uspsbrca.htm


Appendices

List of Figures

1. Transitions Among Four Health States ........................................ 46
2. True Negative Patients: Transition to Death from PCa-free, by Age ............. 47
3. True Positive Patients: Transition to Death from Localized-PCa or Advanced-PCa, by Duration Category and Age ........................................ 48
4. True Positive Patients: Transition to Advanced-PCa from Localized-PCa, by Duration Category and Age ........................................ 49
5. True Positive Patients: Transition to Advanced-PCa from Localized-PCa, by Age Category and Duration ........................................ 50
6. True Positive Patients: Transition to Death from Advanced-PCa for those Initially Diagnosed with Advanced-PCa, by Age Category and Duration .... 51
7. True Positive Patients: Transition to Death from Advanced-PCa for those Initially Diagnosed with Localized-PCa, by Age Category and Duration .......... 52
8. Life Expectancy of 65 year old, by Last Screening Age for PSA Testing and PSA Cut-off Value ........................................ 53

1 Characterization of the social benefit frontier

Let \( V(\cdot) \) denote the value of social benefit under the comparative effectiveness criterion. According to equation (12) and (14), we have the value function as:

\[
V\left( [\psi^p \psi^c]^T, [\varphi^s(\cdot) \varphi^c(\cdot)]^T, [\alpha \overline{\alpha}]^T, [H_t(\cdot) \overline{H}_t(\cdot)]^T, G(\cdot | z), F(z), \gamma \right) = \int \sum_{s=t_1}^{t_f} \gamma^{t-s} [\alpha \overline{\alpha}]^T \cdot \left[ \mathcal{R}_z \left( h_{t_0}, s_{t_0}, \overline{h}_{t_0}, \theta^*(\cdot) \right) \right]_i dG([h_{t_0} \overline{h}_{t_0}]^T | z) dF(z).
\]

This value function represents the frontier of social benefit for the study population.
2 Proof of satisfaction of Condition 2

A confirmatory test that meets Assumptions 3 and 4 has the following confusion matrix:

$$\psi_c = \begin{pmatrix}
\psi_{1,1}^c & \psi_{1,2}^c & \cdots & \psi_{1,n}^c \\
\psi_{2,1}^c & \psi_{2,2}^c & \cdots & \psi_{2,n}^c \\
\vdots & \vdots & \ddots & \vdots \\
\psi_{n,1}^c & \psi_{n,2}^c & \cdots & \psi_{n,n}^c
\end{pmatrix} = \begin{pmatrix}
1 & 0 & \cdots & 0 \\
1 - \lambda_2 & \lambda_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
1 - \lambda_n & 0 & \cdots & \lambda_n
\end{pmatrix}$$

where $0 < \lambda_i \leq 1$, $\forall i = 2, 3, \ldots, n$.

Proof. (Satisfaction of Condition 2) Assumption 4 (i.e., $\psi_{i,j}^c = 0 \ \forall i > 1$ and $j \neq i$) guarantees that a true positive patient’s diagnosed disease is correct. We note that under Assumption 3 (i.e., $\psi_{1,1}^c = 1$), it is impossible to have a false positive patient at any time. That is, $\text{pr}(h^*_t = h_t^1 \text{ and } h^d_t \in \{h^d_2, h^d_3, \ldots, h^d_n\}) = 0$, $\forall t$. Hence, a patient diagnosed with disease is, for sure, a true positive patient. Taken together, a diagnosed patient’s diagnosed disease state is correct.

3 Estimation of the probability of transition from PCa-free to localized-PCa

We choose PCPT data because its enrollees had no evidence of prostate cancer at enrollment and each enrollee was required to have a biopsy test after seven years of follow-up. We define $p$ to be the prevalence of prostate cancer detected at year seven, but require for our analyses the annual transition probability from PCa-free to localized-PCA (denoted $q$). Specifically, $q$ satisfies the following equation

$$p = \sum_{k=1}^{7} \frac{1}{(1 - q)^{k-1}}.$$
we could recover the age-specific probability of transiting from PCa-free to localized-PCa. However, using a constant transition probability is a reasonable first degree of approximation given that we lack access to patient-level PCPT data.

4 Unobserved policy-irrelevant health state and identification

In this appendix, we characterize the impact of non-observability of the policy-irrelevant health state on identification. In what follows, we investigate identification of screening effects for true positive, false negative, and true negative patients in sequence. Finally, we explain how the screening experiment achieves identification in our empirical estimation approach.

As we explained before, estimation without knowledge of diagnosis correctness is extremely challenging due to the intractability of the numerous components that form the likelihood function. As such, our exploration of identification is consistent with our empirical estimation approach described in the text.

Consider a patient at age \( t \) with time-invariant exogenous characteristics, \( z \), who has made a participation decision based on screening policy, \( \theta \). Let \( I_t \) denote information observed by the researcher which includes both patient characteristics and his diagnosis history as well as the true health state that define diagnosis correctness. There exist three types of patients: false negative (FN), true negative (TN), and true positive (TP).

Specifically,

\[
I_t \equiv \begin{cases} 
  z & \text{if FN} \\
  [z, h^*_{t-1}] & \text{if TN} \\
  [h^d_m, t - m, z, h^*_{t-1}] & \text{if TP}
\end{cases}
\]  

(A-17)

where, for true positive patients, both the initial diagnosed disease state \( (h^d_m) \) and diagnosis duration \( (t - m) \) are observed. In addition, the assumption that there exists an indicator of disease progression (based on observed treatments) ensures that \( h^*_{t-1} \) is observed among

\[41\] Assumption 3 rules out the presence of false positive patients.
true positive patients at their post-diagnosis ages. For both false negative and true negative patients, diagnosed disease state \( (h^d_m) \) and diagnosis duration \( (t - m) \) are not applicable. For false negative patients, the most recent policy-relevant health state \( h^*_{t-1} \) cannot be determined. However, for a true negative patient, \( h^*_{t-1} = h^*_1 \).

4.1 True positive patients

By the law of total probability, the patient’s conditional probability of transition from a non-death policy-relevant health state to all-cause death (i.e., \( \{h^*_n \cup \overline{h^*_m}\} \)) given observed patient information and the screening policy is

\[
pr \left( \left\{ h^*_{t-1} \rightarrow \{h^*_n \cup \overline{h^*_m}\} \right\} \mid I_t, \theta \right) \tag{A-18}
\]

\[
= \sum_{\overline{h}^*_{t-1} \in \overline{H}^*} \sum_{\overline{h}^*_n \in \overline{H}^*_n} \sum_{\overline{h}^*_m \in \overline{H}^*_m} pr \left( \left\{ \left[ h^*_{t-1}, \overline{h}^*_{t-1} \right] \rightarrow \left\{ h^*_n \cup \overline{h}^*_m \right\} \right\} \mid I_t, \theta, \overline{h}^*_{t-1} \right) \cdot pr\left( \overline{h}^*_{t-1} \mid I_t, \theta \right)
\]

where,

1. \( h^*_{t-1} \in H^* \setminus h^*_n \) and \( \overline{h}^*_{t-1} \in \overline{H}^* \setminus \overline{h}^*_n \) because the patient is alive at the beginning of age \( t \);
2. \( pr(\overline{h}^*_{t-1} \mid I_t, \theta) \) is the conditional probability that the patient resides at an arbitrary policy-irrelevant health state \( (\overline{h}^*_{t-1}) \) given the observed (by the researcher) information; and
3. our theoretical model indicates that screening policy affects health transition in the policy-relevant dimension only through shifting earliness of diagnosis and, given \( \overline{h}^*_{t-1} \), \( t - m \), and \( h^d_m \) (i.e., \( I_t \), see equation \( \text{(A-17)} \)), screening policy does not affect health transition in the policy-irrelevant dimension. Thus,

\[
pr \left( \left\{ \left[ h^*_{t-1}, \overline{h}^*_{t-1} \right] \rightarrow \left\{ h^*_n \cup \overline{h}^*_m \right\} \right\} \mid I_t, \theta, \overline{h}^*_{t-1} \right) = \sum_{\overline{h}^*_n \in \overline{H}^*_n} \sum_{\overline{h}^*_m \in \overline{H}^*_m} pr \left( \left\{ \left[ h^*_{t-1}, \overline{h}^*_{t-1} \right] \rightarrow \left\{ h^*_n \cup \overline{h}^*_m \right\} \right\} \mid I_t, \overline{h}^*_{t-1} \right).
\]

41
Under the independence assumption between health transition in the policy-relevant dimension and that in the policy-irrelevant dimension, we have

\[
pr\left(\left\{ h_{t-1}^* \rightarrow h_n^* \ | I_t, \overline{h}_{t-1}^* \right\}\right) = pr\left(\left\{ h_{t-1}^* \rightarrow h_n^* \ | I_t, \overline{h}_{t-1}^* \right\}\right) + \frac{\theta}{\theta + 1} \left(1 - pr\left(\left\{ h_{t-1}^* \rightarrow h_n^* \ | I_t, \overline{h}_{t-1}^* \right\}\right)\right)
\]

\[\text{(A-19)}\]

Note, the independence between policy-relevant and policy-irrelevant transitions means

\[
pr\left(\left\{ h_{t-1}^* \rightarrow h_n^* \ | \overline{h}_{t-1}^* \right\}\right) = \frac{\theta}{\theta + 1} \left(1 - pr\left(\left\{ h_{t-1}^* \rightarrow h_n^* \ | \overline{h}_{t-1}^* \right\}\right)\right)
\]

Substituting equation (A-19) into equation (A-18), we obtain the all-cause mortality difference between two arbitrary screening policies (\(\theta\) and \(\theta'\)) conditional upon observed information. That is,

\[
pr\left(\left\{ h_{t-1}^* \rightarrow h_n^* \ | I_t, \theta' \right\}\right) = pr\left(\left\{ h_{t-1}^* \rightarrow h_n^* \ | I_t, \theta \right\}\right) + \frac{\theta}{\theta + 1} \left(1 - pr\left(\left\{ h_{t-1}^* \rightarrow h_n^* \ | I_t, \theta \right\}\right)\right)
\]

\[\text{(A-20)}\]

When the all-cause mortality difference is nonzero, it implies that two patients with identical policy-relevant health states (\(h_{t-1}^*\)), time-invariant characteristics (\(z\)), initial diagnosed disease states (\(h_n^d\)), and diagnosis durations (\(t - m\)) have different all-cause mortalities when the screening policies differ. Such a difference is completely caused by the difference in distribution of policy-irrelevant health (\(\overline{h}_{t-1}^*\)) across the two policies and does not reflect a screening policy effect. When the policy-irrelevant health state is unobserved, the all-cause
mortality difference, if it exists, could be attributed to a screening policy difference, which would lead to an identification concern.

We note that the all-cause mortality difference in equation (A-20) is zero, except when the following inequality holds
\[
pr(h_{t-1}^*|I_t, \theta') \neq pr(h_{t-1}^*|I_t, \theta).
\] (A-21)

To help examine the inequality more closely, we define the following notation. Let \( pr_{t, h^*}(z) \) denote the vector recording the probabilities that a patient resides in one of the \( \pi \) policy-irrelevant health states if he survives to the end of age \( t-1 \). Let \( pr_{t_0, h^*} \) denote the probability that the patient resides in one of the \( \pi \) policy-irrelevant health states at the initial age \( (t_0) \). Let \( P_s(z) \) denote the age-specific \( \pi \) by \( \pi \) policy-irrelevant health transition matrix for the patient with time-invariant characteristics \( z \) at period \( s \).

Recall (1) our model assumption that health transition in the policy-relevant dimension and that in policy-irrelevant dimension are independent and (2) health transition in the policy-irrelevant dimension is a standard Markov process. It follows, by the Chapman-Kolmogorov equation, that the patient’s probabilities of residing in the \( \pi \) policy-irrelevant health states are
\[
pr_{t, h^*}(z) = pr_{t_0, h^*} \prod_{s=t_0}^t P_s(z).
\] (A-22)

Let \([.]_i \) denote the \( i \)th element of vector \([.] \). We obtain that \( \forall i = 1, 2, \ldots, \pi \),
\[
pr(h_{t-1}^* = h_i^*|I_t, \theta) = pr(h_{t-1}^* = h_i^*|h_{m}^{d}, t-m, z, h_{t-1}^*, \theta) = \left[ pr_{t_0, h^*} \prod_{s=t_0}^t P_s(z) \right]_i.
\] (A-23)

Note that the policy-irrelevant health transition, \( P_s(z) \), and screening policy are statistically independent \( (P_s(z) \perp \perp \theta, \forall s = t_0, \ldots, t_J) \), because \( P_s(z) \) is completely governed by the patient’s exogenous characteristics (e.g., race, gender, genetics) in \( z \).\footnote{The screening policy only \textit{indirectly} affects the policy-irrelevant health transition because screening perturbs death due to the disease of interest. Recall, we consider patients who have survived up to the end of period \( t-1 \), so the indirect impact from screening on policy-irrelevant health evolution is \textit{ex post} known to be absent.} Hence, an identification issue may arise due to the identification concern.\footnote{To help examine the inequality more closely, we define the following notation. Let \( pr_{t, h^*}(z) \) denote the vector recording the probabilities that a patient resides in one of the \( \pi \) policy-irrelevant health states if he survives to the end of age \( t-1 \). Let \( pr_{t_0, h^*} \) denote the probability that the patient resides in one of the \( \pi \) policy-irrelevant health states at the initial age \( (t_0) \). Let \( P_s(z) \) denote the age-specific \( \pi \) by \( \pi \) policy-irrelevant health transition matrix for the patient with time-invariant characteristics \( z \) at period \( s \). Recall (1) our model assumption that health transition in the policy-relevant dimension and that in policy-irrelevant dimension are independent and (2) health transition in the policy-irrelevant dimension is a standard Markov process. It follows, by the Chapman-Kolmogorov equation, that the patient’s probabilities of residing in the \( \pi \) policy-irrelevant health states are
\[
pr_{t, h^*}(z) = pr_{t_0, h^*} \prod_{s=t_0}^t P_s(z).
\] (A-22)

Let \([.]_i \) denote the \( i \)th element of vector \([.] \). We obtain that \( \forall i = 1, 2, \ldots, \pi \),
\[
pr(h_{t-1}^* = h_i^*|I_t, \theta) = pr(h_{t-1}^* = h_i^*|h_{m}^{d}, t-m, z, h_{t-1}^*, \theta) = \left[ pr_{t_0, h^*} \prod_{s=t_0}^t P_s(z) \right]_i.
\] (A-23)

Note that the policy-irrelevant health transition, \( P_s(z) \), and screening policy are statistically independent \( (P_s(z) \perp \perp \theta, \forall s = t_0, \ldots, t_J) \), because \( P_s(z) \) is completely governed by the patient’s exogenous characteristics (e.g., race, gender, genetics) in \( z \).\footnote{The screening policy only \textit{indirectly} affects the policy-irrelevant health transition because screening perturbs death due to the disease of interest. Recall, we consider patients who have survived up to the end of period \( t-1 \), so the indirect impact from screening on policy-irrelevant health evolution is \textit{ex post} known to be absent.} Hence, an identification issue may arise due to the identification concern.\footnote{The screening policy only \textit{indirectly} affects the policy-irrelevant health transition because screening perturbs death due to the disease of interest. Recall, we consider patients who have survived up to the end of period \( t-1 \), so the indirect impact from screening on policy-irrelevant health evolution is \textit{ex post} known to be absent.}
concern can arise only if the screening policy is correlated with the initial distribution of policy-irrelevant health states at age \( t_0 \). In summary,

\[
\theta \not\perp \not\perp \text{pr}_{t_0, \theta} \\
\Rightarrow \text{pr}(h_i | I_t, \theta) \neq \text{pr}(h_i | I_t, \theta') \\
\Rightarrow \text{pr} \left( \{h_{t-1}^* \rightarrow \{h_n^* \cup \overline{h}_m^*\} \} | I_t, \theta \right) \neq \text{pr} \left( \{h_{t-1}^* \rightarrow \{h_n^* \cup \overline{h}_m^*\} \} | I_t, \theta' \right).
\]

The heuristic interpretation of this conclusion is if patients under two screening policies are drawn from two different initial policy-irrelevant health state distributions (i.e., \( \theta \not\perp \not\perp \text{pr}_{t_0, \theta} \)), then the differential in health transition between patients under two screening policies, which is caused by their difference in initial policy-irrelevant health states and a screening effect that also impacts health transition (by perturbing earliness of diagnosis and initially diagnosed disease state), cannot be identified separately. This possibility raises an identification concern. Conversely, if patients under two different screening policies are drawn from the same initial policy-irrelevant health state distribution (i.e., \( \theta \perp \perp \text{pr}_{t_0, \theta} \)), then the independence assumption between the policy-relevant health transition and the policy-irrelevant health transition guarantees that, at the beginning of a future period, survivors under the two screening policies have the same policy-irrelevant health distribution. In this case, there is no identification concern because it is safe to attribute the all-cause mortality difference in a period across the two survivor groups to screening policy variation.

### 4.2 False negative and true negative patients

Recall our modeling assumption that the health transition probabilities of a false negative patient who has PCa would not differ if he was newly diagnosed at \( t \) or remained false negative at \( t \). It follows that identification among false negative patients is a special case of that among true positive patients. By imposing \( h_{t-1}^* = h_1^* \), the derivation above for identification among true positive patients still holds because the diagnosed disease state \( h_{d_m}^* \) and diagnosis duration \( (t - m) \), whether applicable or not, does not affect the derivation.
In sum, $\theta \not\perp pr_{t_0, t^*}$ also leads to estimation bias among false negative and true negative patients.

4.3 Screening experiments

A successful randomized screening experiment ensures that patients are drawn from the same initial policy-irrelevant health state distribution across screening arms (i.e., $\theta \perp pr_{t_0, t^*}$). Thus, the identification concern presented above vanishes.
Figure 1: Transitions Among Four Health States
Figure 2: True Negative Patients: Transition to Death from PCa-free, by Age
Figure 3: True Positive Patients: Transition to Death from Localized-PCa or Advanced-PCa, by Duration Category and Age
Figure 4: True Positive Patients: Transition to Advanced-PCa from Localized-PCa, by Duration Category and Age
Figure 5: True Positive Patients: Transition to Advanced-PCa from Localized-PCa, by Age Category and Duration
Figure 6: True Positive Patients: Transition to Death from Advanced-PCa for those Initially Diagnosed with Advanced-PCa, by Age Category and Duration
Figure 7: True Positive Patients: Transition to Death from Advanced-PCa for those Initially Diagnosed with Localized-PCa, by Age Category and Duration
Figure 8: Life Expectancy of 65 year old, by Last Screening Age for PSA Testing and PSA Cut-off Value