Pre-Analysis Plan: Endogenous Screening and Health Behaviors

Draft for Review

November 19, 2025

1 Overview and Objectives

Our paper models the decision to seek low-dose CT (LDCT) lung cancer screening. Because 80% of lung cancers are caused by cigarette smoking, we build a dynamic structural model in which screening and smoking are chosen simultaneously. Screening has the potential to identify early-stage lung cancer, when it is potentially treatable, and it alleviates uncertainty about the true lung cancer state. At the same time, screening is costly in both pecuniary and non-pecuniary terms, and it is not perfect in the sense that false positives may lead to unnecessary care. Smoking generates utility, particularly for those with significant and recent smoking histories, but it also increases the risk for both lung cancer and other chronic conditions. Our goal is to estimate the model and to simulate policy counterfactual scenarios that address both costs and benefits of screening.

To directly estimate the model, we would need longitudinal information on screening and smoking behaviors, subjective beliefs, screening outcomes, and health transitions over many years. Because such data do not exist, our empirical approach combines data from several sources. For our primary data source, we plan to conduct a randomized information experiment that will evaluate how brief, randomized information about *stigma*, *perceived quality*, and *price* affect intentions to seek low-dose CT (LDCT) lung cancer screening within 12 months. Our population of interest are those above the age of 40 who have some cigarette smoking history. Before the information treatments, we will assess eligibility for $0 out-of-pocket (OOP) screening based on USPSTF 2022 guidelines. We will run our survey on Prolific. We intend to report the average treatment effects generated by our experiment and the heterogeneity in these effects by theoretically relevant factors at baseline, including stigma, perceptions of quality, and income. We will then use these treatment effects to calibrate our model. We will complement these treatment effects with additional data drawn from electronic health record data provided by Truveta. In what follows, we present the structural model and our survey design. We also pre-specify the econometric specifications that generate our main survey results, and we discuss how these results will inform the structural model.

2 Structural Model

The goal of the model is to rationalize the observed rate of screening. To do so, we model screening and smoking decisions jointly in an environment with uncertainty about disease state and the quality of screening. We consider two forms of health: lung cancer and other chronic health. Unless lung cancer is advanced such that symptoms are apparent, the state is unobserved by both the individual and the econometrician. Early stage lung cancer, if diagnosed, can potentially be cured, but advanced lung cancer cannot be cured. Other health is a binary variable for the presence of any other chronic health condition, which introduces the notion of competing risk. We present the model in the order of a representative period t.

State Variables

$$\Gamma_t = (\theta_t, b_t, h_t, \phi_t, a_t)$$

- $\theta_t \in \{1, 2, 3\}$: Lung cancer state corresponding to none, early-stage, and late-stage lung cancer. If $\theta = 3$, then θ is observed, otherwise it is unobserved by both the individual and the econometrician.
- b_t: the subjective probability of early-stage lung cancer. If $\theta = 3$, then $b = 0$. Even in the case of a positive screening result in period $t - 1$, the timing of the model is such that θ_t is uncertain.
- h_t: Smoking history.
- $\phi_t \in \{1, 2\}$: Other health state

There is also exogenous age a_t. In this model, age and time are synonymous. We consider a time horizon from age 35 to age 100, at which point death occurs with probability 1.

Actions

Based on their state, an individual simultaneously chooses whether to smoke cigarettes and whether to screen for lung cancer.

$$A_t = (s_t, x_t)$$

- $s_t \in \{0, 1\}$: Smoking decision
- $x_t \in \{0, 1\}$: Screening decision if $\theta_t < 3$

Screening Outcomes

At the time an agent makes smoking and screening decisions, there are two outcomes that are uncertain, both of which are observed if the agent chooses to screen. First, conditional on screening, they receive a signal of early stage lung cancer $z \in \{0, 1\}$. The quality of the test depends on its rates of false positives and false negatives. Define:

$$\Pr(z_t = 1 \mid \theta_t) = \begin{cases} \rho & \text{if } \theta_t = 2 \\ \lambda & \text{if } \theta_t = 1. \end{cases}$$

In this case, ρ is the accuracy of the test (i.e., sensitivity), the probability that the test reveals a true early-stage lung cancer. Thus, $1 - \rho$ is the probability of a false negative. Similarly, λ is the probability of a positive test result in the absence of early-stage lung cancer (i.e., the rate of false positive), and $1 - \lambda$ is the probability of a correctly negative test result. Both ρ and λ have objective average values that we intend to estimate from Truveta data. However, what matters for behavior are an individual's subjective beliefs about $\hat{\rho}$ and $\hat{\lambda}$. We intend to measure average subjective beliefs from the survey.

Treatment is the second outcome. In the event that an agent receives a positive signal, we assume that the agent undergoes additional testing. The subsequent testing, which we assume incurs the same cost as the initial testing, is able to precisely determine the value of $\theta \in \{1, 2\}$. If $\theta_t = 2$, we assume that the agent undergoes immediate treatment y, which affects the θ transition probability at the end of the period. If the screening signal is positive (i.e., if $z_t = 1$) and the agent is free of lung cancer (i.e., $\theta_t = 1$), then we assume that subsequent testing, which is still costly, rules out lung cancer.

Utility Function

Utility is a function of the current state, the actions selected, and the outcomes of screening, if screening is chosen. Specifically, we specify a CRRA function of income net of expenditures on screening, smoking, and cancer treatment, and additively separable terms for smoking, screening, and treatment. Utility from smoking depends on its price p_s and the history of smoking h_t, which captures reinforcement (Becker and Murphy, 1988). The

financial cost of screening is given by p_x, and the disutility of screening is given by γ_4. If the individual must undergo treatment y_t, the disutility of treatment is γ_5.

$$u(\Gamma_t, a_t) = U_\theta + \frac{(W_t - p_s s_t - p_x x_t - p_y y_t)^{1-\gamma_0}}{1-\gamma_0} + (\gamma_1 + \gamma_2 h_t + \gamma_3 h_t^2) \cdot s_t - \gamma_4 x_t - \gamma_5 y_t$$

Utility is health state specific in the sense that we normalize the utility of death. To ensure positive utility when alive, we also specify an additive parameter U_θ that we allow to vary by whether an individual has late-stage cancer ($\theta = 3$) or not ($\theta < 3$). To the extent that $U_{\theta<3} > U_{\theta=3}$, these utility shift terms create dynamic incentives for good health, and they are identified by individual's willingness to engage in preventive care.

Conditional on $\theta_t < 3$, the agent is unaware of the value of θ, and this generates uncertainty about the payoff from screening. Thus, the contribution of current utility to the value function depends on the screening choice. If $\theta_t < 3$ and $x_t = 0$, then the contribution of current utility is simply $u(\Gamma_t, a_t; U_{\theta<3})$. However, if $\theta_t < 3$ and $x_t = 1$, then expected current utility is the expectation over the likelihood of θ and the beliefs about the quality of the screening test:

$$E(u(\Gamma_t, a_t)) = (1-b_t)(1-\lambda)u(\Gamma_t, a_t; \gamma^1, z_t = 0, y_t = 0) + (1-b_t)\lambda u(\Gamma_t, a_t; \gamma^1, z_t = 1, y_t = 0) + \quad (1)$$

$$+(b_t)(1-\rho)u(\Gamma_t, a_t; \gamma^1, z_t = 0, y_t = 0) + (b_t)\rho u(\Gamma_t, a_t; \gamma^1, z_t = 1, y_t = 1) \quad (2)$$

This term is the weighted average of utility over the two possible values of θ (1,2) and the signals from screening. The weights are the subjective probabilities that the agent holds about early stage cancer (b) and the likelihood that the screening test will be accurate (ρ and λ).

Screening Price

The price of screening depends on the agent's state and on exogenous health insurance and location. Because the USPSTF guidelines dictate insurance coverage, we assume that the out-of-pocket price of screening is \$0 if an individual has health insurance and meets the 2022 criteria. We also model geographic access, allowing the full price of screening to reflect the travel and opportunity costs for agents who do not live in an area with screening. Specifically, let

$$p_x = 1[ineligible] * p + 1[distant] * w$$

where eligibility requires that $h_t \geq 20$ and $age \in [50, 79]$. The travel distance, which we capture with a simple binary variable, times the wage rate represent the opportunity costs of screening.

State Evolution

Age and Smoking History
There is no uncertainty about an individual's smoking history. Define the evolution of deterministic state variables smoking history and age:

$$h_{t+1} = h_t + s_t, \quad a_{t+1} = a_t + 1.$$

Cancer Transition
We make simplifying assumptions that the probability of late stage lung cancer conditional no lung cancer is zero, and the probability of no lung cancer conditional on late-stage lung cancer is zero. As such,

- From $\theta_t = 0$:

$$\pi_{12}(h_t, s_t, a_t) = \Pr(\theta_{t+1} = 2 \mid \theta_t = 1, s_t, h_t, a_t) = \text{logit}^{-1}(\alpha_0 + \alpha_1 h_t + \alpha_2 s_t + \alpha_3 a_t) \quad (3)$$

- From $\theta_t = 1$:

$$Recovery: \ \pi_{21}(h_t, s_t, a_t) = \text{logit}^{-1}(\delta_0 + \delta_1 h_t + \delta_2 s_t + \delta_3 a_t + \delta_4 y_t) \tag{4}$$

$$Progression: \ \pi_{23}(h_t, s_t, a_t) = \text{logit}^{-1}(\zeta_0 + \zeta_1 h_t + \zeta_2 s_t + \zeta_3 a_t + \zeta_4 y_t) \tag{5}$$

$$Stasis: \ \pi_{22}(h_t, s_t, a_t) = 1 - \pi_{10} - \pi_{12} \tag{6}$$

Cancer Belief Updating
Beliefs at time t (b_t) reflect subjective beliefs about early stage cancer at time t (θ_t). The goal is to understand how beliefs evolve given possible screening and cancer outcomes. Given our assumption that $\theta = 3$ is an absorbing state, beliefs about early stage lung cancer are no longer relevant to the problem faced by the agent if $\theta = 3$. That is, if $\theta_t = 3$, then belief b_t will not factor in the agent's calculation of future values. If $\theta_t \neq 3$, beliefs update differently given the screening decision and the results of screening. If the agent chooses not to screen, then the posterior belief of early-stage lung cancer is given as:

$$b_{t+1} = b_t + (1 - b_t)\pi_{12},$$

where π_{01} represents the objective transition probability from $\theta = 1$ to $\theta = 2$. This says that agents have rational expectations regarding the average evolution of cancer risk. In the case in which they choose to receive no private signal of information, agent's update beliefs based on this average.

If the person chooses to screen and observes a negative signal $z_t = 0$, then uncertainty about θ_{t+1} stems from both the possibility of a false negative test and the average likelihood of transiting to (or remaining in) early stage cancer. Taking these sources of uncertainty in steps, first, the agent updates beliefs from the negative screening test via Bayes' rule: In this case, beliefs are:

$$\tilde{b} = \frac{(1 - \hat{\rho})b_t}{(1 - \hat{\rho})b_t + \hat{\lambda}(1 - b_t)},$$

and now accounting for the fact that cancer may develop by $t + 1$:

$$b_{t+1} = \tilde{b} + (1 - \tilde{b})\pi_{12}.$$

If the person choose to screen and observes $z_t = 1$, then screening has the effect of resolving uncertainty about period t θ because a positive signal generates confirmatory testing. Thus, updated beliefs about period $t+1$ θ are the rational expectations of early stage lung cancer in $t + 1$ conditional on the known value of θ_t

$$b_{t+1} = \begin{cases} \pi_{12} & \text{if } \theta_t = 1 \\ \pi_{22} & \text{if } \theta_t = 2. \end{cases}$$

Health Transitions
We also assume that the other health state $\phi = 2$ is an absorbing state such that:

$$\Pr(\phi_{t+1} = 2 \mid \phi_t = 2) = 1,$$

and

$$\Pr(\phi_{t+1} = 2 \mid \phi_t = 1, h_t, a_t, s_t) = \text{logit}^{-1}(\psi_0 + \psi_1 h_t + \psi_2 s_t + \psi_3 a_t) \tag{7}$$

Mortality

The model also incorporates mortality, the probability of which is increasing in lung cancer and other health states. Note that mortality is not a function of smoking or smoking history—these terms work through lung cancer and other health.

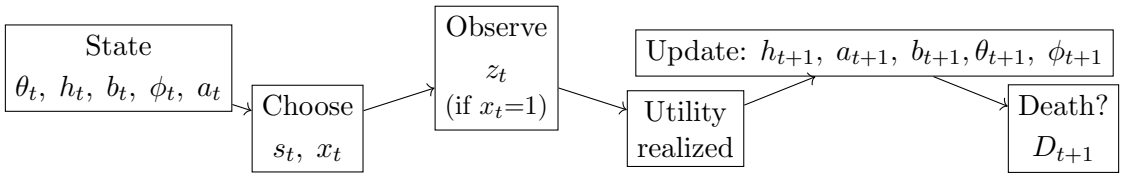
$$\Pr(D_{t+1} = 1 \mid \theta_{t+1}, \phi_{t+1}, a_t) = \text{logit}^{-1}(\omega_0 + \omega_1 \cdot 1[\theta = 1] + \omega_2 \cdot 1[\theta = 2] + \omega_3 \cdot \phi + \omega_4 \cdot a_t) \tag{8}$$

Bellman Equation

Given uncertainty about θ, the value function is the maximum over action space A of the expected value over the perceived distribution of θ.

$$V(\Gamma_t) = \max_{A_t} \Big\{ E(u(\Gamma_t, a_t)) + \beta \mathbb{E}_{\theta,\phi} \left[(1 - P(D_{t+1} = 1)) V(\Gamma_{t+1}) \right] + 1[\theta_t = 2] \left[u(\Gamma_t, a_t, \theta_t = 2) + \beta \mathbb{E} \left[(1 - P(D_{t+1} = 1)) V(\Gamma_{t+1}) \right] \right] \Big\} \tag{9}$$

Model Summary



3 Survey Flow and Experimental Design

We begin by asking standard assessments of risk preferences (Falk et al., 2018) and time preferences (Barsky et al., 1997) that group individual preferences. Next, we assess health insurance status, income, and zip code. Geographic information will be merged at the zip code level to American College of Radiology data lung cancer screening sites.

Next, we assess USPSTF eligibility for subsidized (i.e., $0 OOP) screening. We ask respondents their age, their subjective pack-years of smoking history, and their current smoking status. We also assess beliefs regarding early lung cancer for a generic person that matches their age, smoking history, and overall health. We follow this question with an assessment of how different the respondent feels they are from average. The goal of these questions is to provide estimates for the learning process in the structural model. Next, we ask respondents about their history with and beliefs regarding LDCT lung cancer screening. These beliefs include subjective assessments of false and true positive rates. We also ask two questions about perceived stigma, both regarding support from family and about the signal that screening sends about past smoking behavior. Together with basic demographic information provided by Prolific, our baseline information on respondents includes measures of smoking and screening history, beliefs regarding lung cancer and screening outcome probabilities, and perceptions of stigma related to screening.

The goal of our randomized information experiment is to understand how preferences, beliefs, perceptions regarding screening quality, stigma, and prices shape the demand for lung cancer screening. The flow chart below depicts the structure of our experimental design. We ask all respondents on a 0 to 100 scale how likely they are to seek screening in the next year. We follow this question with a binary question for screening in the next year. Respondents are randomized into one of four framings of the intention to screen questions. All respondents are eligible to be randomized into framings that are neutral, or that emphasizes screening quality or lack of judgment. For those eligible for free screening, the "price" framing emphasizes their eligibility; for those ineligible for free screening, the price framing includes a randomized OOP price.

All respondents are eligible for a neutral framing of the screening question (control):

> On a scale from 0 to 100, where 0 = definitely will not and 100 = definitely will, how likely are you to seek a lung cancer screening (low-dose CT) in the next 12 months?

Similarly, all respondents are eligible for framings that deal with screening quality:

> Out of every 100 people who get today's standard lung cancer screening (low-dose CT), about 12 will have a false alarm (a positive result when no cancer is present). Most of these are cleared with a repeat scan. On a scale from 0 to 100, where 0 = definitely will not and 100 = definitely will, how likely are you to seek a lung cancer screening (low-dose CT) in the next 12 months?

and a positive framing regarding stigma:

> Lung cancer screening is a clinical test like any other. Whether you smoke now or used to, staff are trained to treat every patient with dignity, and screening results are kept private. On a scale from 0 to 100, where 0 = definitely will not and 100 = definitely will, how likely are you to seek a lung cancer screening (low-dose CT) in the next 12 months?

For those who are eligible and who are randomized to the price treatment arm, they receive the following framing:

> Because of your age and smoking history, the United States Preventive Services Task Force recommends annual lung cancer screening. For most people in your situation, insurance covers the test at no cost to you. On a scale from 0 to 100, where 0 = definitely will not and 100 = definitely will, how likely are you to seek a lung cancer screening (low-dose CT) in the next 12 months?

For those ineligible for screening and who are randomized to the price treatment arm, we frame the question as:

> Because of your age and smoking history, you are not recommended for lung cancer screening under the United States Preventive Services Task Force guidelines. If you choose to receive a screening, most people in your situation pay an out-of-pocket price of about $X. On a scale from 0 to 100, where 0 = definitely will not and 100 = definitely will, how likely are you to seek a lung cancer screening (low-dose CT) in the next 12 months?

We randomly vary the value $\$X \in \{100, 300, 500\}$.

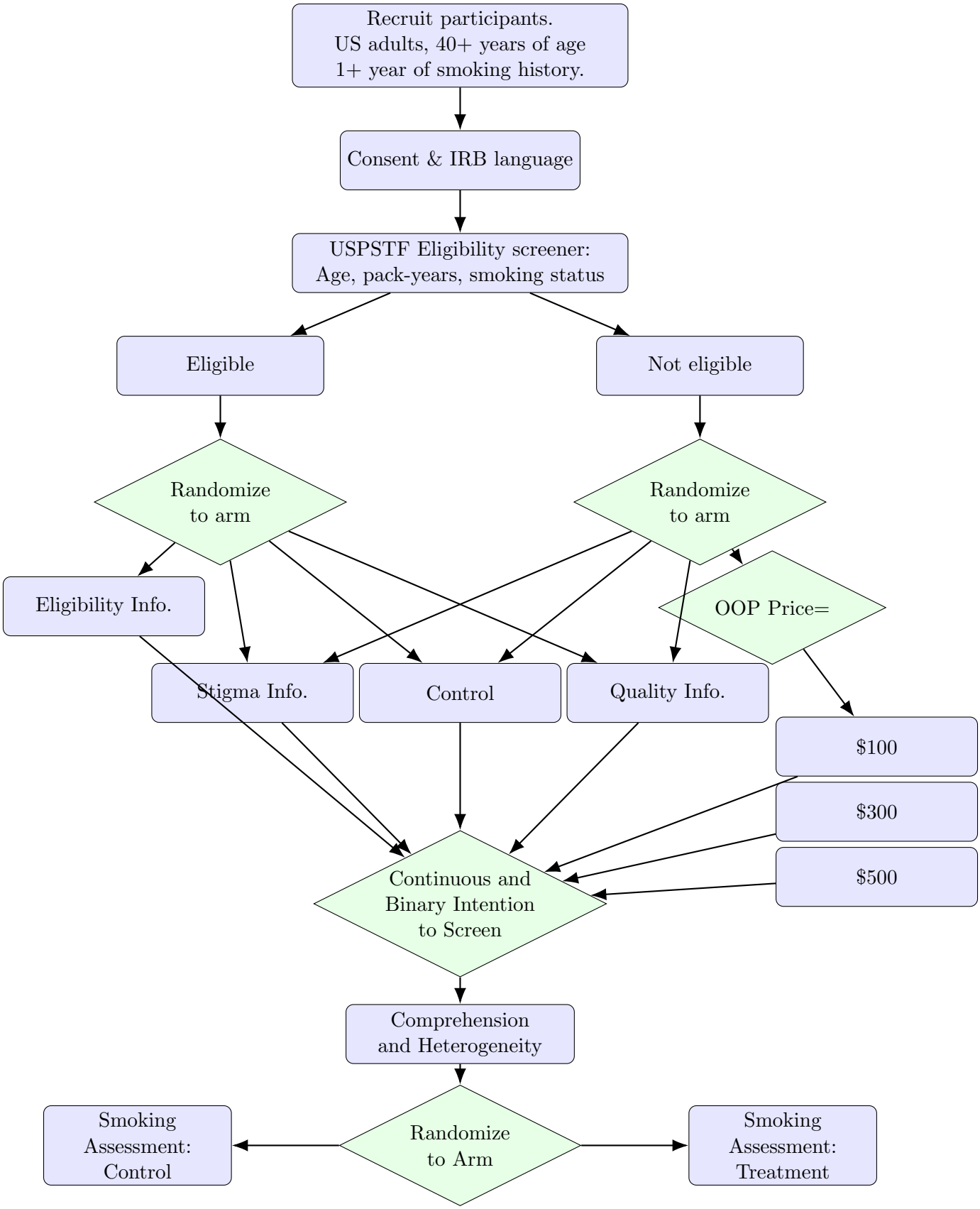
After asking the randomly framed screening intention questions, we ask a series of comprehension questions about the screening that target understanding about the quality of screening, the process of screening (to target stigma) and the cost. We hypothesize that those receiving the respective treatments should do significantly better on these questions. Next, we assess heterogeneity in differences in intention to screen by treatment arm. Specifically, we hypothesize larger effects of the positive stigma framing among those who, at baseline admit to feeling stigma. Our quality framing presents the true false positive rate of screening; for respondents whose subjective assessment of the false positive rate is higher, our treatment represents "good" quality news, and we expect that our quality treatment should encourage screening. Finally, we assess heterogeneity in the price effects by income.

Our survey ends with a randomized framing of a question regarding expecting smoking behavior over the following 12 months. The control arm receives the following neutral framing:

> Thinking about the next 12 months, please indicate which option best reflects your expectations regarding your smoking behavior.

In contrast, to capture how smoking behavior may respond to the results of lung cancer screening, some respondents are presented with the following framing:

> Suppose a recent lung cancer screening revealed no evidence of cancer. Thinking about the next 12 months, please indicate which option best reflects your expectations regarding your smoking behavior:


Recruit participants.
US adults, 40+ years of age
1+ year of smoking history.
Consent & IRB language
USPSTF Eligibility screener:
Age, pack-years, smoking status
Eligible
Not eligible
Randomize
to arm
Randomize
to arm
Eligibility Info.
OOP Price=
Stigma Info.
Control
Quality Info.
$100
$300
$500
Continuous and
Binary Intention
to Screen
Comprehension
and Heterogeneity
Smoking
Assessment:
Control
Randomize
to Arm
Smoking
Assessment:
Treatment

4 Econometric Analysis

Our main regression specification relates intention to screen to indicators for each treatment arm, allowing the price treatment arm effect to vary by eligibility. The equation is:

$$Y_i = \alpha + \beta_S 1\{\text{Subsidy}_i\} + \beta_Q 1\{\text{Quality}_i\} + \beta_P 1\{\text{Price}_i\} + \beta_{PE} 1\{\text{Price}_i\} \cdot 1\{\text{Eligible}_i\} + \mathbf{X}_i'\boldsymbol{\gamma} + \varepsilon_i. \tag{10}$$

Relative to the neutral framing control arm, the interpretation of the stigma and quality parameters is straightforward. The coefficient β_P represents the difference in mean intention to screen between those who are ineligible for free screening and who receive a quoted OOP price relative to the control group. The coefficient β_{PE} represents how receiving information of free screening for eligible respondents affects intention to screen relative to the control group. We intend to present results with and without controls for X, the full set of observable baseline characteristics.

We also estimate the responsiveness to OOP price among ineligible respondents who receive and OPP price. We intend to estimate:

$$Y_i = \gamma + \beta_{300} 1\{Price_i = \$300\} + \beta_{500} 1\{Price_i = \$500\} + \mathbf{X}_i'\boldsymbol{\gamma} + \varepsilon_i, \tag{11}$$

where γ represents the mean intention to screen conditional X among those ineligible respondents who received a quote of \$100, and β_{300} and β_{500} reflect deviations from γ.

Finally, we intent to measure heterogeneity in our main effects in Equation 10 by theoretically relevant baseline characteristics. For example, we hypothesize that the effect of receiving positive experience information will affect those who, at baseline, feel more of a stigma associated with screening. Thus, we estimate:

$$\begin{aligned} Y_i = \alpha &+ \beta_S \, 1\{\text{Subsidy}_i\} + \beta_{SS} \, 1\{\text{Subsidy}_i\} \cdot 1\{\text{High Stigma}_i\} \\ &+ \beta_Q \, 1\{\text{Quality}_i\} + \beta_P \, 1\{\text{Price}_i\} + \beta_{PE} \, 1\{\text{Price}_i\} \cdot 1\{\text{Eligible}_i\} + \mathbf{X}_i'\boldsymbol{\gamma} + \varepsilon_i. \end{aligned} \tag{12}$$

We intend to estimate similar regressions based on baseline information and baseline income. Specifically, we hypothesize that informing individuals that the false positive rate is 12% will have a positive effect on those for whom their baseline assessment was a higher false positive rate. Similarly, we hypothesize that information regarding free screening will have a larger effect on lower income respondents.

4.1 Power Calculations

Our intention is to field the survey to 4,000 respondents, with 25% in each of the four arms. If we assume that 35% of our sample will be USPSTF eligible, that our rich set of baseline characteristics explain 20% of the variation in intention to screen, and that the standard deviation of intention to screen is 25, then the minimum detectable effect for each of our treatment arms is 2.29 percentage points when $\alpha = 0.05$ and the power is set at 0.8.

5 Presentation of Results

Table 1: Baseline Covariate Balance

	Mean	p_1	p_2	p_3
Age				
Female				
Race				
White				
Black				
Other Race				
Less than $25,000				
$25,000-$49,999				
$50,000-$99,999				
$100,000-$199,999				
More than $200,000				
Uninsured				
USPSTF Eligible				
USPSTF Age Eligible				
USPSTF Pack-Years Eligible				
USPSTF Smoking Eligible				
Ever Screened				
Currently Smoke Cigarettes.				
Belief of Early-Stage Lung Cancer				
Lower Expected Risk				
About the Same.				
Higher Expected Risk				
Belief of True Positive				
Belief of False Positive				
High Stigma				
Avoids Information				

Notes: Table 1 presents the overall mean and three p-values. The first p-value is of the F-test that there are no differences in means between the control arm and the three treatment arms quality, stigma, and price. The second p-value is on the F-test of no differences in means between the lowest price treatment ($100) and the the other two ($300 and $500) conditional on being randomized to the price arm and being ineligible for free screening. The final p-value is on the t-test that there is no difference in means between the treatment and control arm of the smoking experiment. $n = XX$

Table 2: Effects of Information on Intent to Screen (ITT)

	Stigma	Quality	Price (Eligible)	Price (Ineligible)
Scale 0-100			*Control Mean =*	
Coef.				
S.E.				
Scale 0-1			*Control Mean =*	
Coef.				
S.E.				
n				
Controls	✓	✓	✓	✓

Notes: Table 2 presents intent-to-treat estimates of from our primary information experiment.

Table 3: Effects of OOP for Ineligible Respondents

	Price = $300	Price = $500
Scale 0-100	*$100 Price Mean =*	
Coef.		
S.E.		
Scale 0-1	*$100 Price Mean =*	
Coef.		
S.E.		
N		
Controls	✓	✓

Notes: Table 3 presents intent-to-treat estimates of from our primary information experiment.

Table 4: Effects of Information on Intent to Screen, Heterogeneity

	Stigma		Quality		Price (Eligible)		Price (Ineligible)	
	Low	High	Bad News	Good News	High Income	Low Income	High Income	Low Income
Scale 0-100					*Control Mean* =			
Coef.								
S.E.								
Scale 0-1					*Control Mean* =			
Coef.								
S.E.								
N								

Notes: Table 4 presents intent-to-treat estimates of from our primary information experiment.

Table 5: Effects of OOP for Ineligible Respondents, Heterogeneity

	Price = $300		Price = $500	
	High Income	Low Income	High Income	Low Income
Scale 0-100	*$100 Price Mean =*			
Coef.				
S.E.				
Scale 0-1	*$100 Price Mean =*			
Coef.				
S.E.				
N				

Notes: Table 5 presents intent-to-treat estimates of from our primary information experiment.

References

Barsky, R. B., Juster, F. T., Kimball, M. S., and Shapiro, M. D. (1997). Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement study. *The Quarterly Journal of Economics*, 112(2):537–579.

Becker, G. S. and Murphy, K. M. (1988). A theory of rational addiction. *Journal of political Economy*, 96(4):675–700.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692.